LREC-COLING 2024

Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)

Workshop Proceedings

Editors
Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre and Alexandre Rademacher

25 May, 2024
Torino, Italia

Proceedings of the LREC-COLING 2024 Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)

ISBN 978-2-493814-20-3
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Preface

Multiword expressions (MWEs) are word combinations that exhibit lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasies (Baldwin and Kim, 2010), such as *by and large, hot dog, pay a visit* and *pull someone's leg*. The notion encompasses closely related phenomena: idioms, compounds, light-verb constructions, phrasal verbs, rhetorical figures, collocations, institutionalized phrases, *etc*. Their behavior is often unpredictable; for example, their meaning often does not result from the direct combination of the meanings of their parts. Given their irregular nature, MWEs often pose complex problems in linguistic modeling (e.g. annotation), NLP tasks (e.g. parsing), and end-user applications (e.g. natural language understanding and MT), hence still representing an open issue for computational linguistics (Constant et al., 2017). This joint workshop also marks the 20th anniversary of the MWE workshop series since 2003 (Bond et al., 2003). The organization of the workshops is sponsored by SIGLEX.[1]

Universal Dependencies (UD; De Marneffe et al., 2021) is a framework for cross-linguistically consistent treebank annotation that has so far been applied to over 100 languages. The framework aims to capture similarities as well as idiosyncrasies among typologically different languages (*e.g.*, morphologically rich languages, pro-drop languages, and languages featuring clitic doubling). The goal of developing UD was not only to support comparative evaluation and cross-lingual learning but also to facilitate multilingual natural language processing and enable comparative linguistic studies. Starting with the first UD workshop in 2017 (de Marneffe et al., 2017), this joint workshop is the 7th edition in the series.

For the current edition, the MWE and UD communities decided to organize a joint event, the MWE-UD workshop which is part of LREC-COLING 2024. This is a timely collaboration because the two communities clearly have overlapping interests. For instance, while UD has several dependency relations that are intended for annotation of MWEs, both annotation guidelines (i.e. is *syntactic irregularity and inflexibility* or *semantic non-compositionality* the leading criterion?) and annotation practice (both across treebanks for a single language and across languages) for MWEs can be improved (Schneider and Zeldes, 2021). For verbal MWEs, the PARSEME corpora for 26 languages provide annotation of MWEs consistent with UD annotation (Savary et al., 2023). Both communities share an interest in developing guidelines, data-sets, and tools that can be applied to a wide range of typologically diverse languages, raising fundamental questions about tokenization, lemmatization, and morphological decomposition of tokens. Proposals for harmonizing annotation practice between what has been achieved in PARSEME and UD and expanding PARSEME annotation to non-verbal MWEs are also central to the recently started UniDive[2] COST action (CA21167). UniDive also co-organizes and sponsors this joint workshop.

We are happy to have received 53 submissions, 29 long, 15 short, and 9 non-archival. 19 long, 7 short, and 9 non-archival contributions were selected for presentation at the workshop, bringing the overall acceptance rate for archival papers to 59%. The distribution over tracks is almost even: 8 of 12 archival submissions were accepted for the UD track, 9 of 16 for the MWE track, and 9 of 16 for the MWE+UD track. One long paper was withdrawn after acceptance.

An important theme in both the UD and MWE community is increasing the number of languages and language families that can be used as the object of study, for instance by making annotated data available in a standard format. The current workshop makes a substantial contribution towards that goal, as it includes contributions to Arabic, Hindi, Old Egyptian, Vedic, Northern Kurdish, Slovene, Dutch, Bavarian, South Asian languages, Turkic languages, Gujarati, Saraiki, Swedish, and more. Another important theme for research on MWEs has been the question

[1] https://siglex.org/

[2] https://unidive.lisn.upsaclay.fr

of to what extent Large Language Models deal adequately with the idiomatic meaning of multi-word expressions. This workshop also includes several contributions that explicitly deal with this question. Apart from these important and cross-disciplinary themes, there are also contributions on UD addressing such issues as assessing and enhancing the value of UD parsing for applications, improved automatic parsing procedures, and the interface between syntax and morphology. Contributions that are primarily concerned with MWEs address a.o. the role of lexical resources, automatic identification of MWEs, the proper annotation of idiomatic meanings in a corpus with fully structured meaning annotation, annotation in parallel corpora, and cross-lingually consistent annotation of MWEs with word senses. Some of these themes re-occur in the contributions that address both UD and MWEs, such as the interplay of lexicon and corpus annotations, the annotation of multiword functional categories, the annotation of light verb constructions, and the use of UD and MWEs in the task of stance detection.

Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre, Alexandre Rademacher.

Acknowledgements

MWE-UD Workshop 2024 has been co-organised and funded by The Special Interest Group for the Lexicon of the Association for Computational Linguistics (ACL-SIGLEX) and the CA21167 COST Action UniDive, supported by COST (European Cooperation in Science and Technology). ACL-SIGLEX funded two (2) participants, while UniDive provided funds for the travel and stay of 29 participants.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors. 2003. *Proceed- ings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.* Association for Computational Linguistics, Sapporo, Japan.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Marie-Catherine de Marneffe, Joakim Nivre, and Sebastian Schuster, editors. 2017. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. Parseme meets universal dependencies: getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9(1).

Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 160–172, Sofia, Bulgaria. Association for Computational Linguistics.

Organizing Committee

Workshop Organizers

Archna Bhatia, Institute for Human and Machine Cognition, USA
Gosse Bouma, Groningen University, Netherlands
A. Seza Doğruöz, Ghent University, Belgium
Kilian Evang, Heinrich Heine University Düsseldorf, Deutschland
Marcos Garcia, University of Santiago de Compostela, Galiza, Spain
Voula Giouli, Institute for Language & Speech Processing, ATHENA RC, Greece
Lifeng Han, University of Manchester, United Kingdom
Joakim Nivre, Uppsala University and Research Institutes of Sweden, Sweden
Alexandre Rademaker, IBM Research, Brazil

Program Committee

Verginica Barbu Mititelu, Cherifa Ben Kehlil, Philippe Blache, Francis Bond, Claire Bonial, Julia Bonn, Tiberiu Boroș, Marie Candito, Giuseppe G. A. Celano, Kenneth Church, Çağrı Çöltekin, Mathieu Constant, Monika Czerepowicka, Daniel Dakota, Miryam de Lhoneux, Marie-Catherine de Marneffe, Valeria de Paiva, Gaël Dias, Kaja Dobrovoljc, Rafael Ehren, Gülşen Eryiğit, Meghdad Farahmand, Christiane Fellbaum, Jennifer Foster, Aggeliki Fotopoulou, Stefan Th. Gries, Bruno Guillaume, Tunga Gungor, Eleonora Guzzi, Laura Kallmeyer, Cvetana Krstev, Timm Lichte, Irina Lobzhanidze, Teresa Lynn, Stella Markantonatou, John P. McCrae, Nurit Melnik, Johanna Monti, Dmitry Nikolaev, Jan Odijk, Petya Osenova, Yannick Parmentier, Agnieszka Patejuk, Pavel Pecina, Ted Pedersen, Prokopis Prokopidis, Manfred Sailer, Tanja Samardžić, Agata Savary, Nathan Schneider, Sabine Schulte im Walde, Sebastian Schuster, Matthew Shardlow, Joaquim Silva, Maria Simi, Ranka Stanković, Ivelina Stoyanova, Stan Szpakowicz, Shiva Taslimipoor, Beata Trawinski, Ashwini Vaidya, Marion Di Marco, Amir Zeldes, Daniel Zeman

Keynote Speakers

Natalia Levshina, Radboud University
Harish Tayyar Madabushi, University of Bath

Organized by

SIGLEX
ACL Special Interest Group

UniDive

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

Funded by
the European Union

Table of Contents

Workshop Program

09:00–09:05 **Welcome**

09:05–09:50 **Keynote Speaker 1**

Every Time We Hire an LLM, the Reasoning Performance of the Linguists Goes Up
Harish Tayyar Madabushi

09:50–10:30 **Session 1: Oral presentations**

09:50–10:10 Assessing BERT's sensitivity to idiomaticity
Li Liu and Francois Lareau

10:10–10:30 Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection
Dylan Phelps, Thomas M. R. Pickard, Maggie Mi, Edward Gow-Smith and Aline Villavicencio

10:30–11:00 **Coffee Break**

11:00–12:00 **Session 2: Poster presentations**

12:00–13:00 **Session 3: Oral presentations**

12:00–12:20 Universal Feature-based Morphological Trees
Federica Gamba, Abishek Stephen and Zdeněk Žabokrtský

12:20–12:40 Light Verb Constructions in Universal Dependencies for South Asian Languages
Abishek Stephen and Daniel Zeman

12:40–13:00 Strategies for the Annotation of Pronominalised Locatives in Turkic Universal Dependency Treebanks
Jonathan Washington, Çağrı Çöltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Gulnura Jumalieva, Aida Kasieva, Aslı Kuzgun, Büşra Marşan and Chihiro Taguchi

13:00–14:00 **Lunch**

14:00–14:45	**Keynote Speaker 2**
	Using Universal Dependencies for testing hypotheses about communicative efficiency Natalia Levshina
14:45–15:00	**Booster Session: Remote presentations**
15:00–16:00	**Session 4: Oral presentations**
15:00–15:20	To Leave No Stone Unturned: Annotating Verbal Idioms in the Parallel Meaning Bank Rafael Ehren, Kilian Evang and Laura Kallmeyer
15:20–15:40	Annotation of Multiword Expressions in the SUK 1.0 Training Corpus of Slovene: Lessons Learned and Future Steps Jaka Čibej, Polona Gantar and Mija Bon
15:40–16:00	Ad Hoc Compounds for Stance Detection Qi Yu, Fabian Schlotterbeck, Henning Wang, Naomi Reichmann, Britta Stolterfoht, Regine Eckardt and Miriam Butt
16:00–16:30	**Coffee Break**
16:30–17:20	**Session 5: Poster presentations**
17:20–17:50	**Community Discussion**
17:50–18:00	**Closing and Awards**

Posters and Remote Presentations

Automatic Manipulation of Training Corpora to Make Parsers Accept Real-world Text
Hiroshi Kanayama, Ran Iwamoto, Masayasu Muraoka, Takuya Ohko and Kohtaroh Miyamoto

Identification and Annotation of Body Part Multiword Expressions in Old Egyptian
Roberto Díaz Hernández

Fitting Fixed Expressions into the UD Mould: Swedish as a Use Case
Lars Ahrenberg

Synthetic-Error Augmented Parsing of Swedish as a Second Language: Experiments with Word Order
Arianna Masciolini, Emilie Francis and Maria Irena Szawerna

The Vedic Compound Dataset
Sven Sellmer and Oliver Hellwig

A Universal Dependencies Treebank for Gujarati
Mayank Jobanputra, Maitrey Mehta and Çağrı Çöltekin

Overcoming Early Saturation on Low-Resource Languages in Multilingual Dependency Parsing
Jiannan Mao, Chenchen Ding, Hour Kaing, Hideki Tanaka, Masao Utiyama and Tadahiro Matsumoto

Part-of-Speech Tagging for Northern Kurdish
Peshmerge Morad, Sina Ahmadi and Lorenzo Gatti

Diachronic Analysis of Multi-word Expression Functional Categories in Scientific English
Diego Alves, Stefania Degaetano-Ortlieb, Elena Schmidt and Elke Teich

Lexicons Gain the Upper Hand in Arabic MWE Identification
Najet Hadj Mohamed, Agata Savary, Cherifa Ben Khelil, Jean-Yves Antoine, Iskandar keskes and Lamia Hadrich-Belguith

Revisiting VMWEs in Hindi: Annotating Layers of Predication
Kanishka Jain and Ashwini Vaidya

Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities
Cvetana Krstev, Ranka Stanković, Aleksandra M. Marković and Teodora Sofija Mihajlov

Combining Grammatical and Relational Approaches. A Hybrid Method for the Identification of Candidate Collocations from Corpora
Damiano Perri, Irene Fioravanti, Osvaldo Gervasi and Stefania Spina

Multiword Expressions between the Corpus and the Lexicon: Universality, Idiosyncrasy, and the Lexicon-Corpus Interface
Verginica Barbu Mititelu, Voula Giouli, Kilian Evang, Daniel Zeman, Petya Osenova, Carole Tiberius, Simon Krek, Stella Markantonatou, Ivelina Stoyanova, Ranka Stanković and Christian Chiarcos

Universal Dependencies for Saraiki
Meesum Alam, Francis Tyers, Emily Hanink and Sandra Kübler

Domain-Weighted Batch Sampling for Neural Dependency Parsing
Jacob Striebel, Daniel Dakota and Sandra Kübler

BERT-based Idiom Identification using Language Translation and Word Cohesion
Arnav Yayavaram, Siddharth Yayavaram, Prajna Devi Upadhyay and Apurba Das

MultiMWE: Building a Multi-lingual Multi-Word Expression (MWE) Parallel Corpora [non-archival]
Lifeng Han, Gareth Jones and Alan Smeaton

AlphaMWE-Arabic: Arabic Edition of Multilingual Parallel Corpora with Multiword Expression Annotations [non-archival]
Najet Hadj Mohamed, Malak Rassem, Lifeng Han and Goran Nenadic

A demonstration of MWE-Finder and MWE-Annotator [non-archival]
Jan Odijk, Martin Kroon, Tijmen Baarda, Ben Bonfil and Sheean Spoel

MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank [non-archival]
Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze and Barbara Plank

Is Less More? Quality, Quantity and Context in Idiom Processing with Natural Language Models [non-archival]
Agne Knietaite, Adam Allsebrook, Anton Minkov, Adam Tomaszewski, Norbert Slinko, Richard Johnson, Thomas M. R. Pickard and Aline Villavicencio

A Corpus of Persian Sentences Annotated with Verbal Multiword Expressions: Development and Guidelines [non-archival]
Vahide Tajalli, Yaldasadat Yarandi, Mahtab sarlak, Mehrnoush Shamsfard and Arezoo Haghbin

Annotating Compositionality Scores for Irish Noun Compounds is Hard Work [non-archival]
Abigail Walsh, Teresa Clifford, Emma Daly, Jane Dunne, Brian Davis and Gearóid Ó Cleircín

Redefining Syntactic and Morphological Tasks for Typologically Diverse Languages [non-archival]
Omer Goldman, Leonie Weissweiler and Reut Tsarfaty

UCxn: Typologically Informed Annotation of Constructions Atop Universal Dependencies [non-archival]
Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft and Nathan Schneider

Keynote Speech

Every Time We Hire an LLM, the Reasoning Performance of the Linguists Goes Up

Harish Tayyar Madabushi
University of Bath

Abstract

Pre-Trained Language Models (PLMs), trained on the cloze-like task of masked language modelling, have demonstrated access to a broad range of linguistic information, including both syntax and semantics. Given their access to both syntax and semantics, coupled with their data-driven foundations, which align with usage-based theories, it is valuable and interesting to examine the constructional information they encode. Early work confirmed that these models have access to a substantial amount of constructional information. However, more recent research focusing on the types of constructions PLMs can accurately interpret, and those they find challenging, suggests that an increase in schematicity correlates with a decline in model proficiency. Crucially, schematicity–the extent to which constructional slots are fixed or allow for a range of elements that satisfy a particular semantic role associated with the slot–correlates to the extent of "reasoning" needed to interpret constructions, a task that poses significant challenges for language models. In this talk, I will begin by reviewing the constructional information encoded in both earlier models and more recent large language models. I will explore how these aspects are intertwined with the models' reasoning abilities and introduce promising new approaches that could integrate theoretical insights from linguistics with practical, data-driven approaches of PLMs.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, page 1. May 25, 2024.

Keynote Speech

Using Universal Dependencies for testing hypotheses about communicative efficiency

Natalia Levshina
Centre for Language Studies
Radboud University, Nijmegen, The Netherlands

Abstract

There is abundant evidence that language structure and use are influenced by language users' tendency to be efficient, trying to minimize the cost-to-benefit ratio of communication (e.g., Hawkins, 2004; Gibson et al., 2019; Levshina, 2022). In my talk I will show how data from corpora annotated with Universal Dependencies can be used for testing hypotheses about the role of communicative efficiency in shaping up language structure and use. The hypotheses are as follows:

1. As discussed by typologists (Sapir, 1921; Sinnemäki, 2008), rigid word order can compensate for lack of formal marking of core arguments. The hypothesis is then that there are positive correlations between the entropy of subject and object in a transitive clause in a corpus and the relative frequency of disambiguating case forms or verb forms. These correlations are expected to minimize the articulation effort involved in the use of argument flags or indices.

2. There is a positive correlation between semantic tightness (Hawkins, 1986), operationalized as Mutual Information between lexemes and syntactic roles, and the relative frequency of verb-final clauses in a corpus. Strong associations between lexemes and roles help to avoid the costs of reanalysis in verb-final languages.

3. There is a negative correlation between the relative frequency of verb-final clauses in the clause and the average number of overt core arguments, which helps to save processing costs required for keeping longer dependencies in mind (cf. Ueno & Polinsky, 2009).

These hypotheses will be tested on corpus data annotated with Universal Dependencies, with the help of mixed-effects models with genealogical and geographic information as random effects.

References

Gibson, Edward, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen & Roger Levy. 2019. How efficiency shapes human language. Trends in Cognitive Science 23(5): 389-407. https://doi.org/10.1016/j.tics.2019.02.003

Hawkins, John A. 1986. A Comparative Typology of English and German: Unifying the Contrasts. London: Croom-Helm.

Hawkins, John A. 2004. Efficiency and Complexity in Grammars. Oxford: Oxford University Press.

Levshina, Natalia. 2022. Communicative Efficiency: Language Structure and Use. Cambridge: Cambridge University Press.

Sapir, Edward. 1921. Language: An Introduction to the Study of Speech. New York: Harcourt.

Sinnemäki, Kaius. 2008. Complexity trade-offs in core argument marking. In: Matti Miestamo, Kaius Sinnemäki and Fred Karlsson (eds.), Language Complexity: Typology, Contact, Change, 67–88. Amsterdam: John Benjamins.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@LREC-COLING-2024, pages 2–3. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

Ueno, Mieko & Maria Polinsky. 2009. Does headedness affect processing? A new look at the VO-OV contrast. Journal of Linguistics 45: 675–710.

Automatic Manipulation of Training Corpora to Make Parsers Accept Real-world Text

Hiroshi Kanayama, Ran Iwamoto, Masayasu Muraoka, Takuya Ohko, Kohtaroh Miyamoto
IBM Research
{hkana@jp., ran.iwamoto1@, mmuraoka@jp., ohkot@jp., kmiya@jp.}ibm.com

Abstract

This paper discusses how to build a practical syntactic analyzer, and addresses the distributional differences between existing corpora and actual documents in applications. As a case study we focus on noun phrases that are not headed by a main verb and sentences without punctuation at the end, which are rare in a number of Universal Dependencies corpora but frequently appear in the real-world use cases of syntactic parsers. We converted the training corpora so that their distribution is closer to that in realistic inputs, and obtained better scores both in general syntax benchmarking and a sentiment detection task, a typical application of dependency analysis.

Keywords: syntax, parsing, Universal Dependencies

1. Introduction

In text processing applications that handle documents such as user reviews and contract documents, accurate syntax parsing is desired for semantic analysis and information extraction. The emerging generative approach also requires the analysis of given utterances to make systems reliable and explainable, such as in retrieval augmented generation (Lewis et al., 2020), and the language models can be improved by incorporating syntactic knowledge (Iwamoto et al., 2023).

Multilingual corpora in Universal Dependencies (UD) (Nivre et al., 2016, 2020) are easily available, and they are used for training and evaluation of syntactic analysis components including tokenizers, part-of-speech (PoS) taggers, and dependency parsers, such as Stanza (Qi et al., 2020), UDPipe (Straka, 2018), spaCy (Honnibal et al., 2020) and Trankit (Nguyen et al., 2021).

However, we found a gap between the standardized UD corpora and the real-world application scenarios. There are many noun phrases (NPs) in reviews such as hotel ones written by a customer as (1), instead of a formal sentence typically with a finite verb in a root node of the syntax tree such as in (2).

(1) A very good hotel close to the park!
(2) I think the hotel is very good because it is close to the park.

Another example of noun phrases is a description in a contract document, such as in (3).

(3) total cost of the services

These noun phrases can appear in many kinds of text documents as the title of a document or section, items in enumeration, a header line of a table, and so on. In addition, in many cases, such strings do not have a period or other punctuation marks at the end.

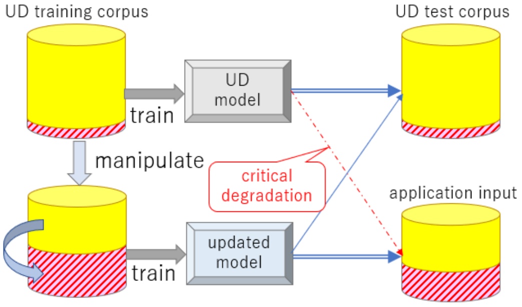
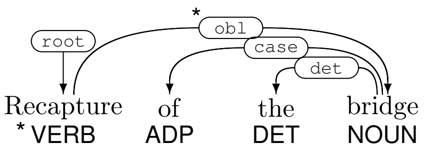


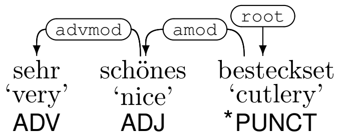
Figure 1: The concept of this paper: an issue of different distribution of text characteristics and its solution by corpus extension.

When we apply a syntax analyzer trained on the UD corpora to such short noun phrases, we often find very wrong analysis results, as exemplified in the output syntax structures of English (4) and German (5). A '*' mark indicates the errors in PoS tags or dependency relations.

(4)



(5)



In (4), the first word "recapture" (which should be NOUN) was incorrectly tagged as VERB as if

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 4–13. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

the input were an interrogative sentence that starts with a verb, and it causes an incorrect dependency relation between "recapture" and "bridge," which should be `nmod` rather than `obl`. In (5), the noun "Besteckset" ('cutlery') was tagged as PUNCT. The writing is not formal because German nouns should start with a capital letter, but the tagging result PUNCT is apparently incorrect. These are actual results by the Stanza parser that achieved very high scores in the UD parsing shared task (Zeman et al., 2018), and we found other taggers and parsers such as UDPipe and spaCy produced similar errors. These errors have already been recognized in the community and discussed in the GitHub issues of those implementations[1].

If most of the contents in the training corpora contain finite verbs in the sentence rather than only noun phrases, it is not surprising that the taggers and parsers trained on such corpora tend to produce incorrect results for the noun phrase inputs such as (4) and (5). Also, we can assume that very unusual tagging results such as in (5) are caused by the training corpus where most sentences end with a period ('.'). Thus, they are problems in the difference between the training corpus and target input to be analyzed.

Figure 1 illustrates the problem that this paper addresses. Normally, the syntax analyzers are trained and evaluated on the UD corpora, but the real-world input documents have different distributions from those of the UD corpora, and the models trained on the UD corpora cause catastrophic errors in applications. Thus, we manipulate the UD corpora to alter distributions in terms of noun phrases and sentence-end punctuation. Although it is impossible to know the general distribution in the real-world inputs, we can make the parser more robust by manipulating the training corpus to reduce the bias in the current UD data.

The contributions of this paper are: (1) to handle the issue regarding noun phrases in addition to punctuation, (2) to provide an algorithm to manipulate training corpora without any manual annotation work, (3) to propose methods to evaluate this work from multiple viewpoints, including the automatic generation of an evaluation data set of noun phrases, and (4) to show the effects of the corpus manipulation in four languages.

Section 2 reviews the related work regarding UD and existing discussions on punctuation and noun phrases. In Section 3, we define the terms used in this paper. Section 4 shows the statistics in different corpora. In Section 5, we propose the algorithm to manipulate training corpora so that the parser can accept real-world inputs, and the effect is shown in Section 6.

2. Related Work

Universal Dependencies (UD) (Nivre et al., 2016) is a worldwide project to provide multilingual syntactic corpora. As of November 2023, 259 treebanks in 148 languages have been released. For all languages, the syntax is represented by dependency trees with 17 PoS tags and 37 dependency labels commonly used for all languages, and each treebank can have language specific extensions. The resources and documentations are available online and incrementally updated.[2] A major shared task of multilingual parsing (Zeman et al., 2018) was held, and a result, UD treebanks is now a de facto standard of multilingual research and many tokenizers and parsers have been trained on them, including a multilingual single parser (Kondratyuk and Straka, 2019).

English Web Treebank (EWT) (Silveira et al., 2014) is one of the most commonly-used treebanks in UD. Originally, it was designed to cover more informal text, such as email and review documents, which was not included in the treebanks of the Wall Street Journal (WSJ). After the emergence of Universal Dependencies, EWT was converted to a UD-style annotation. Thus, EWT contains noisy sentences with typos and abbreviations, and even sentence splitting is tricky (Udagawa et al., 2023), but their work showed that the parsers trained using EWT had a better capability to parse such informal text than the model trained only with WSJ. Due to this historical reason, the EWT corpus functions as an outlier in the experiments in this paper.

The effects of punctuation in a dependency parser have been discussed by Søgaard et al. (2018). They pointed out that dependency parsers, especially neural implementations, are highly sensitive to punctuation in training corpora, and training parsers without punctuation makes the models better. In this paper, we extend the discussion from punctuation to noun phrases, which are more critical in real-world applications. Nivre and Fang (2017) pointed out that punctuation highly affects the benchmarking scores in a number of corpora even if it is not significant in the actual analysis.

The analysis of noun phrase structures have been discussed (Nakov and Hearst, 2005; Vadas and Curran, 2011) but parsing confusion between noun phrases and finite sentences has been less studied. There was a report that a parser specific to noun phrases improved machine translation quality even if the LAS (labeled attachment score) of dependency parsing was not significantly changed (Green, 2011).

[1] An issue of Stanza https://github.com/stanfordnlp/stanza/issues/488 and of spaCy https://github.com/explosion/spaCy/issues/5596.

[2] https://universaldependencies.org/

Corpus synthesis is a powerful method to adapt to specific tasks to enhance a production parser (El-Kurdi et al., 2020) and to broaden the supported languages (Tiedemann and Agic, 2016; Dehouck and Gómez-Rodríguez, 2020) and domains (Li et al., 2019; Jia and Liang, 2016). This paper shares a similar motivation with them but we propose a method to extend training corpora with linguistic knowledge to address specific issues without adding new data sources.

3. Terminology

In this section terms used in this paper are defined.

Unit A text string that is regarded as a single "sentence" in corpora[3]. A unit is also given as an input to a PoS tagger, dependency parser, and their downstream applications, which may be a result of sentence splitting. In this paper we do not call it a "sentence" to distinguish it from the *sentence* defined as follows. All of (1), (2), (3), (4) and (5) in Section 1 can be a unit.

Sentence A unit that is governed by a finite verb, including nominal predicate sentences associated with a copula. A sentence corresponds to a non-terminal symbol 'S' in the phrase structure grammar, though this paper does not discuss its definition from a linguistic viewpoint. Example (2) in Section 1 is a sentence.

Noun Phrase (NP) A syntactic tree or subtree whose head word is a noun or a proper noun, namely, its universal PoS (UPOS) tag is either NOUN or PROPN. An NP also does not have a child node of a copula (where dependency relation label is `cop`). Note that in the content-head structure of UD, the head word of a sentence "She is a teacher." is "teacher" rather than "is" (be-verb).

Noun Phrase Unit (NPU) A unit that forms an NP. Examples include (1), (3), (4) and (5) in Section 1.

Ending punctuation (end-punct) A punctuation mark at the end of a sentence or unit. Here, a punctuation mark is a word that is tagged as PUNCT in the UD corpora. In this paper, we only focus on a period ('.'), an exclamation mark ('!') and a question mark ('?'), which are used in European languages, and discard other PUNCT words like parentheses and quotation marks.

Punctuation Omitted Unit (POU) A unit without ending punctuation. Examples include (3), (4) and (5) in Section 1.

4. Observation of Corpora

To determine how many noun phrase units (NPUs) and punctuation omitted units (POUs) existed in the training corpora and expected input documents, we observed two types of corpora in four languages. One is Universal Dependencies (UD), which is used for the training of various syntax analyzers. Here, we observe the development portion in UD Version 2.13. The other is the review data used for the evaluation of sentiment analysis. We randomly selected 100 sentences[4] of each language version of review data from the SemEval shared task data for aspect-based sentiment analysis (Pontiki et al., 2016) for English, French and Spanish, and Amazon reviews used in another shared task (Ruppenhofer et al., 2014) for German.

Table 1 shows the ratio of the NPUs and POUs in the UD corpora and review documents. Particularly in the UD corpora of French and Spanish, the ratios of NPUs and POUs are very low, that is, the UD corpora tend to consist of formal sentences with finite verbs with ending punctuation marks as their units. The UD English corpus has a relatively higher ratio of NPUs and POUs because there are many informally written documents in EWT corpus as mentioned in Section 2.

The review corpora tend to have many NPUs and POUs, except for the English SemEval data set. There are fewer POUs in SemEval data set (particularly the English one) as expected, that is, most of the units end with a period. The SemEval corpora are supposed to be controlled to have periods for the purpose of extraction of positive or negative expressions with aspects.

As we previously observed, the distribution of syntactic characteristics is very diverse, and those trends highly depend on the formality or cleanliness of the contents of the data set and languages. This shows that it is quite difficult to expect a fixed corpus such as those of UD to represent the distribution of real-world documents that are given to the applications of syntactic analyzers.

5. Corpus Extension

In this section, we propose a method to extend the training corpora for syntactic analyzers, to address the problem of differences in characteristics of corpora discussed in Section 4. Our goal is to build

[3] In the CoNLL-U format used in UD, a unit is represented by a metadata tag '`# text =` '.

[4] Those review data sets do not have syntactic annotation, thus we made manual observation in the limited sentences.

	corpora		NPU ratio (%)		POU ratio (%)	
language	UD	review	UD	review	UD	review
English	EWT	SemEval	23.0	3.0	19.5	1.0
French	GSD	SemEval	3.2	36.0	0.8	3.0
German	GSD	Gestalt	6.1	28.0	1.3	12.0
Spanish	AnCora	SemEval	4.5	25.0	0.8	7.0

Table 1: Ratios of noun phrase units (NPUs) and punctuation omitted units (POUs) in UD and review corpora of four languages.

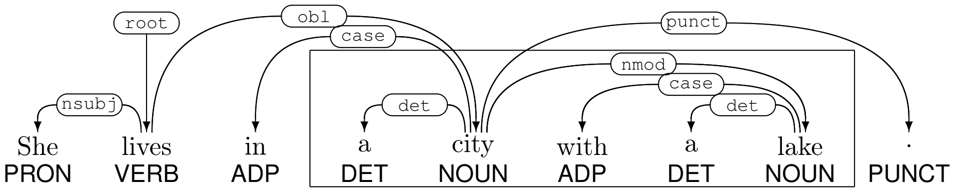


Figure 2: Extraction example of an NP (indicated as a box) from a sentence.

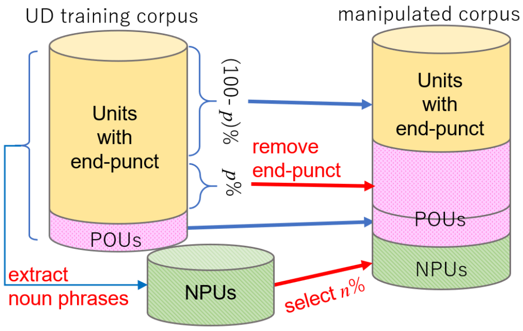


Figure 3: Training corpus extension.

models useful in real applications by reducing the bias in the training model to the UD corpora as illustrated in Figure 3.

5.1. Removal of punctuation

We assume that the incorrectly assigned PUNCT tag to a noun in (5) is caused by the PoS tagging model trained on the corpus where most of last word is tagged as PUNCT. A desirable model is robust to the existence of punctuation, that is, the result should be consistent with or without an end-punct.

A straightforward solution to the problem of the bias to the training corpora is to reduce end-punct at a certain ratio p, in other words, to add POUs, and then to retrain models. Most end-puncts do not have any child nodes in the dependency tree, and thus, it is quite easy to remove an end-punct from a unit, maintaining the validity of the tree[5].

5.2. Addition of noun phrases

In addition to sentences headed by a finite verb, a training corpus should contain noun phrases as units, to handle similar inputs in real applications. To make such a corpus, we add NPUs by extracting noun phrase subtrees from the original corpus in the following manner.

- Identify nouns (a word tagged as NOUN or PROPN) in the original dependency trees.
- Find noun phrases, selecting nouns whose subtree headed by the noun consists of more than three sequential words[6].
- Exclude a preposition and punctuation that should not be a part of a noun phrase. This treatment is needed because the syntactic structure in UD is designed in a content-head manner, and thus, a number of function words are included in a subtree of noun phrases. Functional words that attach to the head of the noun phrase with a dependency label `case` or `punct` are removed from the noun phrase. In the case shown in Figure 2, the preposition “in” is excluded from the noun phrase headed by “city” because it attaches to “city” with `case` relation.
- Pool the noun phrases extracted in this way, and randomly select a number of them in a given ratio (n) to add them to the training corpus, keeping all of the original units in the corpus.

[5] As an exception, the UD_English-EWT training corpus contained a unit in which a conjunction “and” attached to a period at the end of the sentence. In such a case, we did not apply the modification of punctuation removal.

[6] If the children or descendant nodes have a gap due to non-projectivity, such noun phrases are ignored.

In Section 6, we will show the effects of corpora conversion by changing the ratio of punctuation removal and noun phrase addition.

6. Experiments

6.1. Data for evaluation

We will evaluate the syntax analyzer trained on the extended corpora in three ways using three different data sets in four languages: English, French, German and Spanish.

6.1.1. Noun phrase

We evaluate the robustness of the syntax analyzers to the input strings of noun phrases, as a unit test of our approach. For this purpose, we automatically generated the test set of noun phrases in the following procedure.

- Obtain section titles[7] of Wikipedia articles of four languages
- Extract section titles that consist of three or more words
- Exclude those that contain special characters such as numbers, symbols, quotation and punctuation marks
- Exclude those containing non-canonical upper/lower cases (e.g. “RNAb”, “AIESEC”)
- Exclude those that were judged as different languages from that of Wikipedia
- For English, French and Spanish, change the initial character of each word into lower case
- Remove duplication
- Diversify the first word so that there are no more than three entries that share the first word. This is to reduce frequent patterns such as “List of XX”
- Randomly select 1,500 entities for each language

This process almost perfectly extracts noun phrases in each language, and by definition, the last word is not punctuation. Table 2 shows examples in four languages.

In the experiments in this section, we will apply PoS taggers and dependency parsers to these data to calculate the following two scores:

Wrong punctuation The number of cases where the last word is tagged as PUNCT or its dependency label is `punct`. A lower number is better.

NP detection The ratio of the dependency trees of which the root node is tagged as NOUN. A higher ratio is better.

6.1.2. Universal Dependencies

We use the UD corpora for the intrinsic evaluation of dependency parsers. The F1 score of LAS is used as a representative evaluation metric. In our experiments, we extend the *train* and *dev* portions of the UD corpora with the methods presented in Section 5, and the *test* portion for evaluation is not changed. This means the distributions of units are different between the test and training corpora. As a result, the LAS score on the UD test corpus will be theoretically decreased, and thus, minimizing the downgrade of the LAS score indicates the success of our approach.

6.1.3. Sentiment detection

We also conduct an extrinsic evaluation using multilingual sentiment detection (Kanayama and Iwamoto, 2020; Iwamoto et al., 2021) as an application of dependency parsing. For the evaluation, we used sentiment analysis data sets that were observed in Section 4. Those data sets for four languages were derived from shared tasks (Pontiki et al., 2016; Ruppenhofer et al., 2014) and all of them are customer's review data in a domain per language (restaurant for English, French and Spanish, cutlery for German). Each of them contains 500 units, and the annotations were simplified so that each unit has a unit-level polarity flag (either positive or negative) as shown in Table 3.

Similarly to the previous work on multilingual sentiment detection (Kanayama and Iwamoto, 2020), we calculated precision and recall as metrics. Precision depends on the quality of the sentiment lexicon and handling of syntax phenomena such as negation. Recall is related to the coverage of the sentiment lexicon and accuracy in detection of the root node in dependency analysis. The experiments in this paper have few factors that change the precision of sentiment detection, and thus, we focus on recall as it is affected by syntactic structures related to noun phrases.

6.2. Parser retraining

We applied the two kinds of conversion described in Section 5 to the training portions of the UD corpora in four languages (German-GSD, French-GSD, Spanish-AnCora and English-EWT), and retrained models of the Stanza version 1.1.1 (Qi et al.,

[7] Note that they are different from the titles of articles because the majority of article titles are proper nouns, and they are not appropriate to test our method because names are not confused with sentences, and movie titles are hard to determine the desirable annotation (e.g. “Gone with the Wind”).

English	all passenger trains
	cobordism of manifolds with additional structure
French	ponts sur d'autres cours d'eau ('bridges over other waterways')
	instance vérité et dignité ('Truth and Dignity Commission')
German	Meine Daten und ich ('My data and I')
	Mangelnde wissenschaftliche Grundlage ('Lack of scientific basis')
Spanish	recopilatorios y discos especiales ('compilations and special discs')
	contenido de agua en el suelo ('water content in the soil')

Table 2: Examples of noun phrases in the Wikipedia section title data set.

English	This has got to be one of the most overrated restaurants in Brooklyn.	Negative
	Best Pastrami I ever had and great portion without being ridiculous.	Positive
French	Aucune commande de dessert n'a été prise après une demie heure d'attente à la fin de le plat. ('No dessert order was taken after half an hour wait at the end of the dish.')	Negative
	Petit restaurant à le décor soigné, à les tables bien mises. ('Small restaurant with neat decoration, well-set tables')	Positive
German	Die Griffe sind schön geformt, die Messer liegen angenehm in der Hand und sind scharf. ('The handles are beautifully shaped, the knives are comfortable to hold and sharp.')	Positive
	Rostflecken nach Spülmaschine ('Rust spots on dishwasher')	Negative
Spanish	El servicio es muy bueno y la calidad de la comida al mismo nivel. ('The service is very good and the quality of the food at the same level.')	Positive
	Un restaurante al que no pienso volver. ('A restaurant which I don't want to come back to')	Negative

Table 3: Examples of sentiment polarity data. The second example of each language is a noun phrase.

2020) with the extended training corpora. For all languages, we retrained PoS tagging models (`pos`) and dependency parsing models (`depparse`) with maximum iteration of 5,000 times[8], and other models for tokenization (`tokenize`), multi-word tokens (`mwt`) and lemmatization (`lemma`) were not changed from the default ones.

We tested various ratios for the removal of punctuation (p) and addition of noun phrases (n). $p = 0$, $n = 0$ means the original UD corpus as it is, and thus, it is the baseline for each language. We evaluated two scores using the noun phrase data sets described in Section 6.1: number of incorrect punctuation and ratio of NP detection. We also evaluated the LAS score using the UD test corpus, and the recall of sentiment detection using the review corpus.

Stanza's retraining process is randomized and the resultant models are not deterministic, and thus, we conducted 10-times retraining on the baseline settings ($p = 0$ and $n = 0$) to report the average and standard deviation of each score.

6.3. Results

Tables 4, 5, 6 and 7 show the results of all metrics for German, French, Spanish, and English, respectively. The top row ($p = 0, n = 0$) shows the baseline scores with the model trained on the original corpus. The next section (remove punct) shows the effects of reducing end-punct by p, and the last section (add NP) reports the scores by adding NPs to the training corpus varying n, including combination of both modification with p and n.

In the baseline models of German, French and Spanish, there were 3.2 to 4.2$\%$ of catastrophic punctuation errors. Removing end-puncts effectively reduced such errors, even with a small ratio of p. By setting $p = 20\%$, such errors were completely avoided in the four languages.

However, just removing punctuation did not improve the scores of other metrics, although there are a number of settings that improved NP detection in French and Spanish. Also, the changes of LAS and sentiment recall were marginal. The large decrease of LAS scores for $p = 100\%$ (3 points decrease in German and French) is as expected because $p = 100\%$ means all end-puncts were removed from the training corpora, and the punctuation marks that remain in the test corpora are difficult to handle with the model trained by the training corpora without any end-puncts.

[8] Setting `max_steps=5000`, one tenth of the default setting. This is to reduce training time with small sacrifice of accuracy.

	(%)		Section title				UD		Sentiment	
	p	n	Wrong punct (↓)		NP detection (↑)		LAS (↑)		Recall (↑)	
baseline	0	0	3.2	±1.75	97.4	±0.16	79.68	±0.25	52.1	±1.0
remove punct	10	0	**0**	+	97.3		79.85		**53.2**	+
	20	0	**0**	+	97.1		78.98		51.0	−
	50	0	**0**	+	97.3		79.73		50.4	−
	100	0	**0**	+	97.4		76.78	−	49.6	−
add NP	0	10	**0**	+	**98.1**	+	79.59		52.9	
	10	10	**0**	+	**97.8**	+	79.87		**54.6**	+
	20	10	**0**	+	97.5		**80.20**	+	52.9	
	0	20	**0**	+	**97.7**	+	79.64		**53.8**	+
	0	50	**0**	+	**98.1**	+	79.23	−	51.5	
	50	50	**0**	+	**97.9**	+	79.70		52.4	
	0	100	**0**	+	**98.4**	+	79.60		51.5	

Table 4: Results of syntax analysis and sentiment detection in German using the models trained on the extended UD corpora with p punctuation removal and n noun phrase addition. In percent except for the number of incorrect punctuation marks. The top row ($p = 0$, $n = 0$) shows the baseline scores with the original corpus, with the average score of 10 trials and standard deviation. In other rows, a bold number with a + mark indicates that the score is significantly better than the baseline with a difference higher than the standard deviation. A − mark indicates the score is worse against the baseline.

			Section title				UD		Sentiment	
	p	n	Wrong punct (↓)		NP detection (↑)		LAS (↑)		Recall (↑)	
baseline	0	0	4.2	±0.55	91.4	±0.55	87.14	±0.18	43.0	±0.5
remove punct	10	0	**1**	+	**92.5**	+	87.01		42.6	
	20	0	**0**	+	90.6		87.31	−	43.2	
	50	0	**0**	+	91.1		**87.57**	+	**43.6**	+
	100	0	**0**	+	**92.3**	+	84.57	−	42.0	−
add NP	0	10	**3**	+	**93.2**	+	**87.33**	+	42.0	−
	10	10	**0**	+	**93.2**	+	87.09		43.0	
	20	10	**0**	+	**92.9**	+	87.19		42.4	−
	0	20	**0**	+	**93.2**	+	87.25		**44.0**	+
	0	50	**0**	+	**94.4**	+	86.76	−	42.6	
	50	50	**0**	+	**93.6**	+	87.02		42.8	
	0	100	**0**	+	**95.5**	+	86.37	−	**43.6**	+

Table 5: French results. See the caption of Table 4 for details.

			Section title				UD		Sentiment	
	p	n	Wrong punct (↓)		NP detection (↑)		LAS (↑)		Recall (↑)	
baseline	0	0	4.1	±2.90	91.5	±0.68	87.58	±0.16	37.5	±0.6
remove punct	10	0	**0**	+	91.3		87.63		37.8	
	20	0	**0**	+	**93.5**	+	87.28	−	36.4	−
	50	0	**0**	+	91.0		87.52		38.0	
	100	0	**0**	+	91.9		86.83		37.8	
add NP	0	10	**1**	+	**93.1**	+	**88.21**	+	37.2	
	10	10	**0**	+	**92.7**	+	87.67		36.8	
	20	10	**0**	+	**93.1**	+	87.28	−	37.6	
	0	20	**1**	+	**92.8**	+	**88.02**	+	37.8	
	0	50	**1**	+	**94.2**	+	87.37	−	**38.4**	+
	50	50	**0**	+	**94.4**	+	87.52		38.0	
	0	100	**0**	+	**94.7**	+	87.59		**38.2**	+

Table 6: Spanish results. See the caption of Table 4 for details.

The addition of noun phrases had larger impacts in all metrics. When the noun phrases were added ($p = 0$, $n > 0$), NP detection ratio was improved in all four languages, and it was consistently increased with n. Considering that the noun phrases extracted from the UD corpora and those in the Wikipedia section data are independent, we can say that the addition of noun phrases to the training corpora has a positive impact on the analysis of noun phrase inputs generally. There were cases

			Section title				UD		Sentiment	
	p	n	Wrong punct (↓)		NP detection (↑)		LAS (↑)		Recall (↑)	
baseline	0	0	0.7	±0.67	91.6	±0.63	83.84	±0.14	48.9	±0.9
remove punct	10	0	**0**	+	91.7		83.81		47.6	−
	20	0	**0**	+	91.1		**84.06**	+	49.2	
	50	0	2		91.4		**84.03**	+	49.4	
	100	0	**0**	+	90.1	−	83.46	−	49.2	
add NP	0	10	1		**93.9**	+	**84.09**	+	49.0	
	10	10	**0**	+	**94.2**	+	83.71		49.0	
	20	10	1		**93.7**	+	83.96		49.6	
	0	20	**0**	+	**94.6**	+	83.91		49.6	
	0	50	**0**	+	**95.3**	+	83.88		48.6	
	50	50	**0**	+	**95.4**	+	83.75		47.6	−
	0	100	**0**	+	**95.3**	+	**84.00**	+	48.8	

Table 7: English results. See the caption of Table 4 for details.

that were not detected as nouns even for $n = 100\%$, but a number of remaining errors were due to automatic noun phrase extraction from Wikipedia section titles.

The addition of NPs reduced the punctuation errors as well, even without explicit removal of punctuation (e.g. $p = 0$ cases). This is because the noun phrases added to the corpus did not have end-puncts, and thus, it helped models avoid bias to corpora consisting of POUs.

Although these treatments for noun phrase inputs obviously made positive impacts to the Wikipedia section title data, there is a potential risk of damage to the existing benchmarking. In the results of the LAS score in the UD test corpora, the decrease in general dependency parsing performance was observed in a number of cases with high ratios of p and n, but in most of cases, LAS scores were equal to or better than the baseline settings.

Because our motivation in this work is to build a robust parser for real-world applications, an extrinsic evaluation should be a main focus. In French, German and Spanish, recall scores in sentiment detection were increased with a moderate ratio of end-punct removal or NP addition, even though the optimal ratio of p and n varies by languages.

In English, the sentiment detection was not improved from the baseline. These results can be supported by the observation in Section 4: UD_English-EWT data contains NPUs and POUs with higher ratios compared to other corpora, and the English version of SemEval data was highly controlled with formal sentences without NPUs and POUs, and thus, our approach to corpus expansion did not work for this settings, but it is notable that negative impacts were limited as well.

7. Conclusion

This paper presented methods to make robust PoS taggers and dependency parsers to inputs for real-world applications by reducing the discrepancy of the ratios of noun phrases and punctuation omitted units between the training corpora and expected input documents. In addition to the removal of punctuation, which has been attempted to build more consistent models, we added noun phrases to the training corpus by automatically extracting noun phrases from existing annotations using syntactic operations. The experimental results showed that retraining on the extended training corpora made positive impacts on all three experiments simultaneously; a unit test for noun phrases, intrinsic evaluation of the dependency parser, and extrinsic evaluation of it on sentiment detection. The selection of the optimal values in the corpus expansion (ratios of punctuation removal and noun phrase addition) is our future work.

In this paper we handled multiple European languages where the definition of noun phrases and punctuation is relatively easy. In other languages, the structure of noun phrases is more diverse and complicated, and thus, more linguistic discussion and empirical studies will be needed. We applied the proposed technique to the UD corpora, but this can be integrated with the corpus augmented method using raw corpora (El-Kurdi et al., 2020), so that more applicable syntax analyzers can be developed.

The results of our experiments suggest that the current UD corpora are not perfect to train models for practical syntactic analyzers, and that it is important to know the characteristics of corpora and input documents to analyze, and to adjust the corpora to generate better models not just for the benchmarking on UD, but also for the practical use cases.

8. Bibliographical References

Mathieu Dehouck and Carlos Gómez-Rodríguez. 2020. Data augmentation via subtree swapping for dependency parsing of low-resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3818–3830, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yousef El-Kurdi, Hiroshi Kanayama, Efsun Sarioglu Kayi, Vittorio Castelli, Todd Ward, and Radu Florian. 2020. Scalable cross-lingual treebank synthesis for improved production dependency parsers. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 172–178, Online. International Committee on Computational Linguistics.

Nathan Green. 2011. Effects of noun phrase bracketing in dependency parsing and machine translation. In *Proceedings of the ACL 2011 Student Session*, pages 69–74, Portland, OR, USA. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Ran Iwamoto, Hiroshi Kanayama, Alexandre Rademaker, and Takuya Ohko. 2021. A Universal Dependencies corpora maintenance methodology using downstream application. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 23–31, Online. Association for Computational Linguistics.

Ran Iwamoto, Issei Yoshida, Hiroshi Kanayama, Takuya Ohko, and Masayasu Muraoka. 2023. Incorporating syntactic knowledge into pre-trained language model using optimization for overcoming catastrophic forgetting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10981–10993, Singapore. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2016. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Berlin, Germany. Association for Computational Linguistics.

Hiroshi Kanayama and Ran Iwamoto. 2020. How Universal are Universal Dependencies? Exploiting Syntax for Multilingual Clause-level Sentiment Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4063–4073.

Daniel Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Zuchao Li, Junru Zhou, Hai Zhao, and Rui Wang. 2019. Cross-domain transfer learning for dependency parsing. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 835–844. Springer.

Preslav Nakov and Marti Hearst. 2005. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 17–24, Ann Arbor, Michigan. Association for Computational Linguistics.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4034–4043, Marseille, France. European Language Resources Association.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Josef Ruppenhofer, Roman Klinger, Julia Maria Struß, Jonathan Sonntag, and Michael Wiegand. 2014. IGGSA shared tasks on German sentiment analysis (GESTALT). In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 164–173, Hildesheim, Germany. Universität Heidelberg.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2897–2904.

Anders Søgaard, Miryam de Lhoneux, and Isabelle Augenstein. 2018. Nightmare at test time: How punctuation prevents parsers from generalizing. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 25–29.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Jörg Tiedemann and Zeljko Agic. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research*, 55:209–248.

Takuma Udagawa, Hiroshi Kanayama, and Issei Yoshida. 2023. Sentence identification with BOS and EOS label combinations. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 343–358, Dubrovnik, Croatia. Association for Computational Linguistics.

David Vadas and James R. Curran. 2011. Parsing noun phrases in the Penn Treebank. *Computational Linguistics*, 37(4):753–809.

Daniel Zeman, Filip Ginter, Jan Hajič, Joakim Nivre, Martin Popel, and Milan Straka. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium.

Assessing BERT's sensitivity to idiomaticity

Li Liu, François Lareau
OLST, Université de Montréal
Montréal, Canada
{li.liu.2, francois.lareau}@umontreal.ca

Abstract

BERT-like language models have been demonstrated to capture the idiomatic meaning of multiword expressions. Linguists have also shown that idioms have varying degrees of idiomaticity. In this paper, we assess CamemBERT's sensitivity to the degree of idiomaticity within idioms, as well as the dependency of this sensitivity on part of speech and idiom length. We used a demasking task on tokens from 3,127 idioms and 22,551 tokens corresponding to simple lexemes taken from the French Lexical Network (LN-fr), and observed that CamemBERT performs distinctly on tokens embedded within idioms compared to simple ones. When demasking tokens within idioms, the model is not proficient in discerning their level of idiomaticity. Moreover, regardless of idiomaticity, CamemBERT excels at handling function words. The length of idioms also impacts CamemBERT's performance to a certain extent. The last two observations partly explain the difference between the model's performance on idioms versus simple lexemes. We conclude that the model treats idioms differently from simple lexemes, but that it does not capture the difference in compositionality between subclasses of idioms.

Keywords: phraseology, idioms, idiomaticity, multiword expressions (MWEs), language models

1. Introduction

Multiword expressions (MWEs) are characterized by the constrained selection of their components and their partial or complete lack of compositionality (Mel'čuk, 2023). In this paper, we focus on idioms, a prominent category of MWEs known for their non-compositional nature which have long presented a significant challenge for natural language processing (NLP) (Sag et al., 2002; Baldwin and Kim, 2010; Constant et al., 2017).

Idioms cannot be understood simply by the regular combination of the meanings of their components, e.g., *spill the beans* means 'disclose a secret', which cannot be obtained from 'spill'+'beans'. However, while all idioms violate compositionality, some idioms do include the meaning of some or even all of their components, making them more or less semantically transparent. Hence, compositionality in idioms falls on a continuum. According to the degree of inclusion of the meaning of their components, Mel'čuk (2023) classifies idioms into **weak idioms**, which include the meaning of all of their components along with some arbitrary meaning, as in (1), **semi-idioms**, which include the meaning of some but not all of their components along with some arbitrary meaning, as in (2), and **strong idioms**, which are completely non-compositional, as in (3). This is illustrated below with French idioms.

(1) étoile de mer
star of sea
'starfish' = 'star-shaped marine animal'

(2) fruit de mer
fruit of sea
'seafood' = 'food that comes from the sea'

(3) noyer le poisson
drown the fish
'obfuscate things'

The contextualized language model BERT (Devlin et al., 2019), pre-trained on extensive linguistic data, has been widely used and has shown exceptional performance across diverse NLP tasks. Given the high degree of conventionality of idioms (Calzolari et al., 2002), there is a natural expectation for BERT to be good at handling them. Indeed, Tan and Jiang (2021) have validated the model's ability to distinguish between the literal and idiomatic usage of potential idiomatic expressions. Nedumpozhimana and Kelleher (2021) have shown that BERT incorporates information from idioms and their surrounding context to process them. Tian et al. (2023) have demonstrated that BERT-like language models represent idioms differently from their literal counterparts at both sentence and word levels, with words in idioms receiving less attention than words in non-idiomatic contexts. Clearly, BERT has a strong ability at handling idioms. However, one question remains: **is BERT sensitive to the degree of idiomaticity of idioms?**

Our hypotheses are that:

1. CamemBERT should be better at predicting tokens within idioms as opposed to simple lexemes, because tokens within idioms are more strongly constrained.
2. Tokens within idioms with higher idiomaticity should be more likely to be accurately predicted compared to tokens within idioms with lower idiomaticity.

As far as we know, there has been limited research into this question. The closest research was

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 14–23. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

by Garcia et al. (2021b), who conducted a series of probing tasks to examine whether and to what extent vector space models, including BERT, can appropriately represent idiomaticity in noun compounds (NCs) in English and Portuguese. However, their results do not address the following questions: Does BERT distinguish different degrees of idiomaticity in NCs and other types of idioms? What kinds of tokens within an idiom are more predictable? Does the length of an idiom influence BERT's ability to predict tokens within it?

In this paper, we try to answer these questions by focusing on semantic idiomaticity in French idioms. We took our data from the French Lexical Network (LN-fr), a handcrafted lexical resource containing 3,127 idioms, 22,551 simple lexemes, and 47,395 contextual sentences for these entries. Our experiment used CamemBERT-base (Martin et al., 2020), a pre-trained BERT-derived model for French, in a demasking task on both simple lexemes and tokens embedded within idioms from our dataset.

We compared the prediction results of simple lexemes and tokens within idioms to observe performance differences under different conditions, thereby inferring the model's representation of different level of idiomaticity. Moreover, we analyzed the effect of token part of speech (POS) and idiom length on performance.

2. Related work

In recent years, attention has been focused on detecting and representing idiomaticity. Handling a MWE within a context requires first recognizing its non-compositional nature and then accurately conveying its idiomatic meaning in this context. Currently, the primary approach involves generating embeddings for components of the MWE and then merging them using diverse composition functions to construct a comprehensive representation of the MWE. Ultimately, the idiomaticity can be evaluated by computing the cosine similarity between the merged vector and the vector representing the expression (Cordeiro et al., 2019).

To represent idiomatic meaning in MWEs, recent approaches typically utilize contextualized language models. Among these models, Shwartz and Dagan (2019) found that BERT outperforms other contextualized models implemented in classifiers for creating embeddings in tasks related to lexical composition. However, Nandakumar et al. (2019) and Garcia et al. (2021a,b) indicated that pre-trained contextual models cannot effectively encode idiomaticity in MWEs. In comparison, static models like word2vec perform better (King and Cook, 2018; Nandakumar et al., 2018, 2019; Cordeiro et al., 2019; Sarlak et al., 2023). Nevertheless, supervised approaches leveraging contextualized models tend to outshine in tasks specific to certain languages and types of MWEs with ample resources, as these models offer representations that encode linguistic features and contextual cues (Fakharian and Cook, 2021).

Idiomaticity has also become a topic of recent NLP conference tasks. For instance, SemEval-2022 task 2 (Tayyar Madabushi et al., 2022) focuses on idiomaticity detection and sentence embedding containing multilingual MWEs. Results of these tasks show that the models got better performance with available training data. Although the best-performing methods are based on deep neural models independent of the linguistic features of MWEs, mixed approaches are generally believed to be worth exploring. Additionally, the PARSEME shared task on automatic identification of verbal MWEs (Ramisch et al., 2020), particularly with the *Seen*2020 system (Pasquer et al., 2020), underscores the significance of incorporating linguistic features in MWE-related tasks as well.

In our study, we focused on evaluating language models' sensitivity to idiomaticity. For this, we observed the contextualised model CamemBERT's performance in a classic fill-mask task with simple and idiomatic tokens in French.

3. Experiment

3.1. Data

We extracted our data form LN-fr v3 (Polguère, 2009; Lux-Pogodalla and Polguère, 2011; Polguère, 2014; ATILF, 2023), released in October 2023. It is an extensive, openly accessible lexical resource constructed manually following the methodological principles of explanatory combinatorial lexicology (ECL), the lexicological branch of Meaning-Text Theory (MTT) (Mel'čuk and Polguère, 1987; Mel'čuk et al., 1995; Apresjan, 2000). Every entry in LN-fr is a disambiguated lexical unit, i.e., either a simple lexeme or an idiom with a specific meaning, and each idiom is classified as a weak idiom, a semi-idiom or a strong idiom (see §1). Since our study follows MTT's definition and classification of idioms, and because LN-fr contains explicit information about the idiomaticity level of idioms, it suited our purpose very well.

Each lexical unit has a POS tag, and that of an idiom is determined by its internal syntactic head rather than its function within a sentence (Mel'čuk, 2006). For instance, *bien sûr* ('of course', lit. 'well sure'), because its head *sûr* is an adjective, is described as an adjectival idiom despite functioning as an adverb in sentences. There are a total of 11 POS tags for idioms in our dataset (see Table 2).[1]

[1] Interjective idioms are expressions that function as

Lexical unit	Idiomaticity	POS	Examples
pomme	simple lexeme	N	À la fin du repas, on a parfois droit à un petit morceau de brie et, en guise de dessert, selon la saison, des pommes, des noix, quelques fraises écrasées avec du sucre qu'on étale sur une tartine.
pomme de terre	weak idiom	N Prep N	Ils prenaient une demi-heure à midi pour manger un œuf sur le plat, une pomme de terre, du fromage blanc. Pierre avait peine à soulever des sacs de pommes de terre de 40 kg, quant à moi je fis un véritable travail de garçon de ferme.

Table 1: Sample data from LN-fr

Idiom type	Example	Count
Nominal	*coup de soleil* lit. 'blow of sun' 'sunburn'	1579
Prepositional	*à propos* lit. 'at purpose' 'by the way'	730
Verbal	*faire la tête* lit. 'make the head' 'sulk'	619
Conjunctive	*quand même* lit. 'when even' 'anyway'	93
Adjectival	*bien sûr* lit. 'well sure' 'of course'	42
Phrasal	*Un ange passe.* lit. 'An angel passes.' 'awkward silence'	27
Adverbial	*pas mal* lit. 'not bad' 'quite good'	23
Propositional	*qui se respecte* lit. 'who respects oneself' 'self-respecting'	5
Numeral	*un à un* lit. 'one to one' 'one by one'	5
Pronominal	*ici et là* lit. 'here and there' 'here and there'	2
Interjective	*Tonnerre de Dieu!* lit. 'thunder of God' 'Good heavens!'	2
Total		**3127**

Table 2: Idiom types in the dataset

The POS of the tokens that are embedded within an idiom is not annotated directly in LN-fr, but one can retrieve it from the idiom's syntactic pattern, which is a string representing a sequence of POS tags. For example, *pomme de terre* ('potato', lit. 'apple of ground'), has the pattern `N Prep N`, so we know that the first and last tokens are nouns and the second is a preposition. We extracted from these patterns the POS tags for most of the embedded tokens. As some idioms did not have a syntactic pattern, we were not able to automatically retrieve the POS for their embedded tokens, which represent about 3.8% of all the tokens in our dataset; these tokens were not included in our second analysis (§4.2).

Each lexical unit has one or more lexicographic examples taken from corpora. These examples have been meticulously selected by lexicographers to reflect the authentic usage of a lexical unit. They aim to showcase various constructions that are possible for the lexical unit, to illustrate its usage and its syntactic and semantic selection (Lux-Pogodalla, 2014). Moreover, the annotation explicitly gives the position, within each sentence, of the tokens that belong to the lexical unit at hand. Note that a lexical unit may appear more than once in the same example; we counted those separately (which is why we have more tokens than examples even for simple lexemes in Table 3). We had in our dataset a total of 47,395 such sentences, with an average of 1.5 examples per idiom and 2 per simple lexeme, each sentence having around 38 tokens on average.

Finally, we counted the length in tokens of each lexical unit. For simple lexemes the length is 1; for idioms, we segmented by spaces and punctuations.

In total, we extracted from LN-fr 25,678 lexical units: 3,127 idioms and 22,551 simple lexemes. Table 3 breaks down these numbers. Compared to the NCs dataset used by Garcia et al. (2021b) covering 9,220 naturalistic and neutral sentences for 280 NCs in English and 180 NCs in Portuguese, our dataset encompasses a broader spectrum of idioms and a larger quantity of contexts.

Our dataset is available at `https://github.com/liliulng/idiomaticity-dataset`.

3.2. Methodology

Our experiment consists in taking the sentences associated with a lexical unit in LN-fr and masking, one at a time in the case of idioms, the tokens that correspond to that lexical unit. We then submit these sentences to CamemBERT for demasking. The model predicts the masked token and provides

independent sentences, like interjections such as *Wow!*

Type	Lexical units	Examples	Tokens
Simple lexeme	22551	42849	45563
Idiom	3127	4546	13529
Weak idiom	592	916	2425
Semi-idiom	589	899	2408
Strong idiom	1946	2731	8696
Total	**25678**	**47395**	**59092**

Table 3: Quantitative overview of our dataset

a list of candidates, each with a softmax score reflecting the model's confidence in it being the missing token. We record the confidence score returned by the model for the correct answer (the masked token) and note whether the correct answer was ranked as the first candidate (R1). This is illustrated in Table 4. The R1 candidate is the model's best guess and should be viewed as its "answer". Its score tends to be close to 1 (indeed, the model is optimized for this), but sometimes it can be lower, which reflects the model's confidence in its answer (or lack thereof). We want to take this into account, so if the masked token is guessed at rank 1, we note its score, and we will refer to it as "score@R1" in the rest of this paper.

We did not fine-tune the model because we aimed to evaluate the model's ability to learn idioms without being explicitly trained for it. We used the model as-is with its default parameters.

CamemBERT, as a contextualised model, provides predictions of a masked token based on its context. In our case, the contexts are the sentences retrieved from LN-fr that illustrate the usage of simple lexemes and idioms. Because we mask each token within idioms one by one, the other tokens inside a given idiom are visible and are part of the context. Nedumpozhimana and Kelleher (2021) suggested that BERT's ability to understand an idiom primarily relies on the idiom itself, so context inside idioms is crucial for CamemBERT to predict masked idiomatic tokens.

We utilized the model's tokenizer to segment the tokens, guaranteeing that our tokenization was consistent with the model's vocabulary. In cases where a token was segmented into subtokens, such as the token *tigers* being tokenized into *_tiger* and *s*, we conducted the masking experiment for each subtoken and calculated the product of all subtokens' confidence scores as the confidence score for that token. Furthermore, if the model correctly predicted each subtoken, we marked the whole token as correctly predicted as well.

We analysed the distribution of confidence scores of tokens, scores at rank 1 (scores@R1) and the percentage of correct predictions for masked tokens belonging to simple lexemes and idioms with different idiomaticity degrees, in order to determine how much the model's prediction is related to masked token's contextual idiomaticity degree. We further conducted statistical tests to validate the conclusions drawn from our observations.

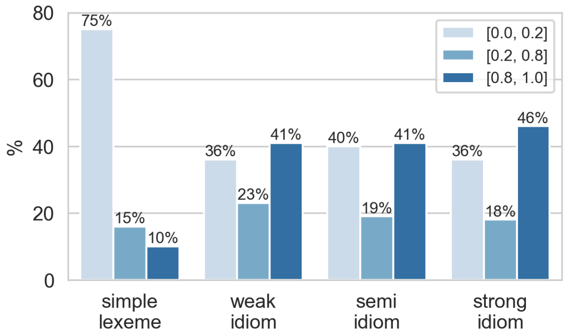


Figure 1: Score distribution

4. Results and Discussion

In this section, we explore the impact of idiomaticity, POS, and idiom length on the model's performance. We examine the confidence scores, scores@R1, and the probability of achieving correct predictions token (expressed as a percentage of R1). When analyzing the scores and scores@R1, we take into account the median and mean for tokens across various categories. These are represented, respectively, by a thick line and a triangle in our figures. When there is a notable difference between them, our focus will be on the median.

4.1. Does CamemBERT distinguish different levels of idiomaticity?

Figure 1 shows that 75% of non-idiomatic tokens score below 0.2, with only 10% achieving a high score above 0.8. Conversely, over 40% of idiomatic tokens are predicted with scores exceeding 0.8, highlighting the model's significant challenge in predicting non-idiomatic tokens. Regarding idiomatic tokens, the model's confidence scores for correct answers often fall into polarized categories of high or low scores. However, discerning between varying levels of idiomaticity remains difficult, as indicated by similar score distributions across the three types of idioms.

The Kruskal-Wallis test proved the significant difference between the confidence score distribution for tokens corresponding to simple lexemes and that of tokens belonging to idioms ($p < 0.01$, $\eta^2 = 0.15$). There is no significant difference between scores for tokens in the three types of idioms ($p < 0.01$, but with negligible effect size $\eta^2 < 0.01$).

When comparing the mean and median confidence scores (Figure 2), we further notice a

Lexical unit	Token	POS	Sentence	Score	R1
pomme	pommes	N	À la fin du repas, on a parfois droit à un petit morceau de brie et, en guise de dessert, selon la saison, des **<mask>**, des noix, quelques fraises écrasées avec du sucre qu'on étale sur une tartine.	0.10	F
pomme de terre	pomme	N	Ils prenaient une demi-heure à midi pour manger un œuf sur le plat, une **<mask>** de terre, du fromage blanc.	0.99	T
	de	Prep	Ils prenaient une demi-heure à midi pour manger un œuf sur le plat, une pomme **<mask>** terre, du fromage blanc.	0.99	T
	terre	N	Ils prenaient une demi-heure à midi pour manger un œuf sur le plat, une pomme de **<mask>**, du fromage blanc.	0.99	T

Table 4: Sample fill-mask inputs and results

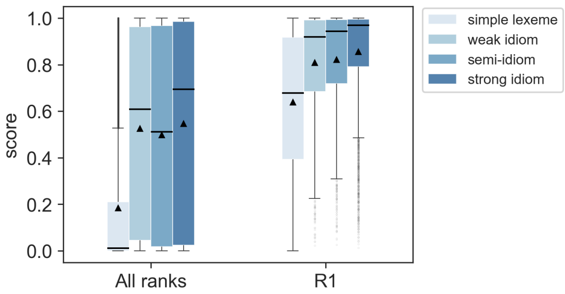


Figure 2: Score given to the masked token at all ranks and at R1

	All	Content	Function
Simple lexemes	25	24	50
Weak idioms	62	55	86
Semi-idioms	58	48	83
Strong idioms	62	49	81

Table 5: Percentage of correctly predicted tokens for content and function tokens

significant difference between idiomatic and non-idiomatic tokens. Idiomatic tokens consistently exhibit higher median and mean scores, typically around 0.5 or above. Still, there is no substantial distinction among the three classes of idioms, as tokens within each category demonstrate fairly similar median and mean scores. However, it is worth noting that score@R1, which represents the model's overall confidence in its predictions, tends to correlate positively with the degree of idiomaticity, which aligns with our previous hypothesis. Additionally, tokens within strong idioms consistently receive the highest median and mean scores, compared to other idiomatic tokens.

R1 predictions: The model correctly guesses the masked token around 60% of the time for tokens within idioms, compared to only 25% for simple lexemes. There is no significant difference between the three types of idioms: 62% for weak idioms, 58% for semi-idioms and 62% for strong idioms. This reveals again the model's higher capacity in predicting tokens within idioms than simple lexemes.

Statistical analysis: We calculated the Spearman's ρ correlation to unveil the dependence of the model's prediction results (confidence scores and scores@R1) on tokens' idiomaticity levels.

Between the free versus idiomatic nature of masked tokens and their prediction results, there is a moderately positive correlation that confirms the model's capability to distinguish tokens on these two general levels, with $p < 0.01$, Spearman's $\rho = 0.36$ for scores and $p < 0.01$, Spearman's $\rho = 0.39$ for score@R1. Specifically for all the four levels of idiomaticity (simple lexeme, weak idiom, semi-idiom, strong idiom), this moderately positive correlation still exists between idiomaticity levels and the prediction results (with $p < 0.01$, Spearman's $\rho = 0.36$ for confidence scores and $p < 0.01$, Spearman's $\rho = 0.38$ for score@R1). As observed in Figure 2, no significant correlation is found between the scores and the three subtypes of idioms ($p = 0.04$, $\rho = 0.02$).

This indicates again that, in general, the model is unable to differentiate between varying levels of idiomaticity within idioms, although it effectively distinguishes between free and idiomatic tokens. A chi-squared test between the idiomaticity levels and correct prediction aligns with this conclusion: $p < 0.01$ and a moderate effect size Cramér's $V = 0.3$ for all idiomaticity levels and the generally free and idiomatic levels, but $p < 0.01$, Cramér's $V = 0.03$ between the three types of idioms.

4.2. What kinds of tokens are more predictable within idioms?

We aimed to pinpoint which kinds of tokens present greater predictive challenge and to understand how this might contribute to the observations above. To accomplish this, we broke down our data by the POS of both free and idiomatic tokens. This data was readily available in LN-fr, which distinguishes

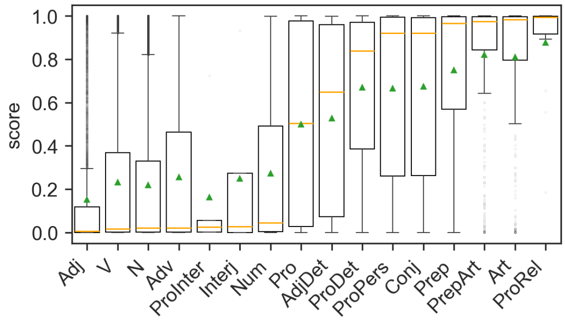


Figure 3: Score by token POS

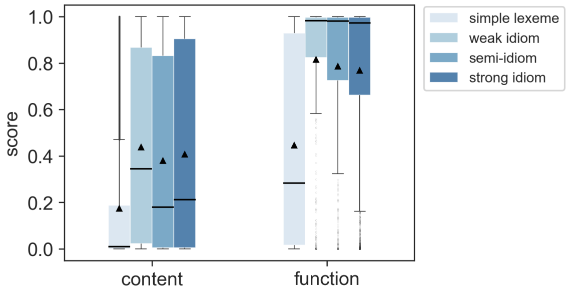


Figure 4: Scores for content and function words

a total of 16 POS tags (distinct from the 11 for idioms listed in Table 2) that can be divided into two categories: content and function tokens. Content tokens represent 94% of the tokens in our dataset and include nouns (N), verbs (V), adjectives (Adj), adverbs (Adv), numerals (Num), interrogative pronouns (ProInter) and interjections (Interj). Function tokens represent the other 6% and include pronouns (Pro), prepositions (Prep), articles (Art), preposition-article amalgams (PrepArt), conjunctions (Conj), personal pronouns (ProPer), pronominal determiners (ProDet), adjectival determiners (AdjDet) and relative pronouns (ProRel). Three of these categories had very low counts, namely Interj (4 occurrences), ProInter (14) and ProRel (5), so the scores reported here for those categories are to be taken with a grain of salt (this explains why the mean is outside of the box for ProInter).

As Figure 3 shows, the median and mean scores for all function tokens are notably higher than those for content tokens, exceeding 0.5. Conversely, the median and mean confidence scores for content tokens are low, with mean scores below 0.3 and median scores below 0.1. This suggests that overall, disregarding idiomaticity, the model excels in predicting function tokens. The score@R1 exhibits the same trend, hence we omit the graph here.

R1 predictions: 82% of function tokens were correctly predicted, against only 28% of content tokens.

Statistical analysis: Spearman's ρ test demonstrated a moderately positive correlation between predictions and type of POS (content or function token): with $p < 0.01$, $\rho = 0.31$ for confidence scores and $p < 0.01$, $\rho = 0.39$ for score@R1. The chi-squared test also detected a certain level of dependence between the correct prediction of tokens and their POS status ($p < 0.01$, Cramér's $V = 0.3$)

These results are not surprising, because function words belong to closed classes, thus there are far fewer options for the model to choose from. However, given the model's adeptness at managing function tokens, we wondered if this could explain its better performance on idioms. Indeed, there is a stark contrast between the distribution of content and function tokens in simple lexemes versus idioms: function tokens comprise only 0.5% of the simple lexemes, while they account for 28.6% of the tokens within idioms. This is because idioms are phrases, so they often contain function words, especially in French, where compounds are much less common than in some other languages such as English or Chinese. Hence, could this imbalance account for the elevated median and mean scores observed for tokens within idioms reported in Figure 2?

We analyzed separately the confidence scores of content and function tokens with varying degrees of idiomaticity. As shown in Figure 4, regarding content tokens, the median and mean scores of idiomatic tokens generally fall below 0.5 but still remain significantly higher than those for simple lexemes. Similarly, there is no substantial disparity in scores among tokens in different types of idioms for content tokens. As for function tokens, those within idioms receive higher confidence scores overall, with mean scores surpassing 0.7 and median scores nearing 1. The variance among different types of idioms is minimal. Conversely, scores for simple function tokens are notably lower than those for idiomatic function tokens, below 0.5. Thus, regardless of the degree of idiomaticity, the model's prediction of function tokens consistently outperforms that of content tokens. As for content tokens, the model's prediction of tokens within idioms surpasses that of simple lexemes, and its prediction ability for tokens within idioms with varying degrees of idiomaticity remains stable. This corresponds to our previous conclusion in the first analysis (see §4.1).

R1 predictions: The percentages of correct predictions for content tokens across various levels of idiomaticity further support our findings (see Table 5). Specifically, more than 50% of the content tokens within idioms were correctly predicted, compared to only 24% for simple content tokens. In addition, while roughly half of simple function tokens were correctly predicted, this figure exceeded 80% for idiomatic function tokens.

Statistical analysis: We conducted the same statistical analysis for prediction results across idiomaticity levels for content and function tokens separately. Spearman's ρ correlation between idiomaticity levels and confidence scores or score@R1 always yielded $p < 0.01$ but with no significant ρ values. The chi-squared test showed only modest dependence between correct prediction and idiomaticity level (either considering all four levels or only free versus idiomatic), for both content and function tokens: $p < 0.01$, Cramér's $V = 0.2$. There is no clear dependence between prediction results and the three idiomaticity levels across idiom subtypes ($p < 0.01$, Cramér's $V = 0.05$). Thus, the moderate correlation between idiomaticity levels and correct prediction observed in the first analysis no longer exists when we separate content and function tokens. This suggests that the variation in prediction performance of the model between free and idiomatic tokens may actually be at least partly due to the differing proportions of content and function words in these tokens.

No specific POS within content or function tokens appears to significantly influence the model's performance. The primary types of content words include nouns, verbs, and adjectives. In both simple lexemes and idioms, nouns comprise most of the words, accounting for approximately 61% in simple lexemes, 75% in weak idioms, 78% in semi-idioms, and 66% in strong idioms. Verbs represent a similar portion in simple lexemes (21%) and strong idioms (16%), while they only make up 4% and 6% in weak idioms and semi-idioms. There is no significant difference in the proportion of adjectives across simple lexemes and idioms, ranging from approximately 12% to 18%. Confidence scores for nouns, verbs, and adjectives do not show significant differences. As for function tokens, pronouns (59%), conjunctions (24%), and personal pronouns (10%) are the primary function token types in simple lexemes, while prepositions constitute the main portion of the function tokens in idioms, comprising 74% in weak idioms, 78% in semi-idioms, and 60% in strong idioms. Additionally, preposition-articles are the second major type, accounting for 17%, 14%, and 13% respectively in the aforementioned subtypes of idioms. Notably, the proportion of articles in strong idioms is higher at 15% compared to weak and semi-idioms (2% and 4%).

To sum up, function words tend to be accurately predicted by the model in all types of expressions regardless of the level of idiomaticity, because they belong to closed classes with a small number of members. In free context, their predictability arises from governing syntactic relations and sentence coherence. Meanwhile, within idioms, they contribute to idiomaticity by maintaining the structural integrity and idiomatic meaning of the expression.

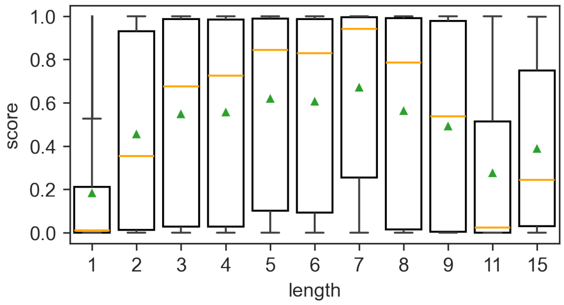


Figure 5: Scores by lexical unit length

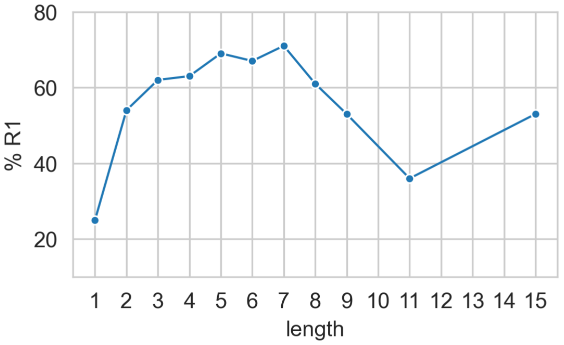


Figure 6: Percentage of R1 by lexical unit length

4.3. Is CamemBERT sensitive to the length of idioms?

When a token in an idiom is masked, CamemBERT utilizes contextual information to predict the masked one, and that context includes the remaining tokens in the idiom. Therefore, the more tokens an idiom contains, the more context it provides. Consequently, does CamemBERT achieve better prediction results for tokens within longer idioms? In our dataset, 99% of idioms comprise 7 tokens or fewer, whereas longer idioms amount to only 171 occurrences, representing only 1% of the idioms. Most idioms, specifically 91% of weak idioms, 87% of semi-idioms, and 59% of strong idioms, consist of 2 or 3 tokens. Additionally, a small proportion (8% of semi-idioms and 20% of strong idioms) extend to 4 tokens, while another 10% of strong idioms span 5 tokens. No statistically significant relation is found between the level of idiomaticity and the length of idioms.

We compared the score and score@R1 for tokens in lexical units of varying lengths. Here again, the results for score@R1 are not different, so we only present the results for confidence scores in Figure 5. They suggest that as the length of lexical units increases, both the mean and median confidence scores tend to rise (we disregard the drop for lengths over 7 tokens, which we attribute to the scarcity of data in that range).

R1 predictions: Similarly, as shown in Figure 6, when the length of idioms is 7 tokens or fewer, there is a generally increasing trend between idiom length and the percentage of correct predictions.

Statistical analysis: With $p < 0.01$, the Spearman's ρ coefficient between lexical unit length and scores is 0.36, while it is 0.4 for score@R1, suggesting a moderate positive correlation. Similarly, correct prediction displays a moderate positive association with idiom length in the chi-squared test ($p < 0.01$, Cramér's $V = 0.32$). These findings suggest that the length of idioms significantly impacts CamemBERT's prediction of idiomatic tokens. The model evidently demonstrates sensitivity to the length of idioms when interpreting tokens within them.

Due to the small proportion (1%) of idioms with lengths exceeding 7 tokens, and despite their proportion of correct predictions not aligning with the general trend, their impact has been disregarded in our analysis.

5. Conclusion

We aimed to assess CamemBERT's ability to capture varying degrees of idiomaticity within idioms. We measured this by comparing the model's off-the-shelf performance on fill-mask tasks with tokens pertaining either to simple lexemes or idioms, further distinguishing three levels of idiomaticity among idioms: weak idioms, semi-idioms and strong idioms. We collected 59,092 tokens with illustrative examples from LN-fr, including 45,563 simple lexemes and 13,529 idiomatic tokens from more than 3,000 idioms.

In §1, we posited two hypotheses:

1. CamemBERT should be better at predicting tokens within idioms as opposed to simple lexemes.
2. Tokens within idioms with higher idiomaticity should be more likely to be accurately predicted.

Our main observations are:

1. The model is significantly better at predicting tokens that belong to an idiom as opposed to simple lexemes.
2. It is not sensitive to varying levels of idiomaticity among subtypes of idioms.
3. It exhibits a heightened performance in predicting function words, regardless of idiomaticity.
4. There is a positive correlation between idiom length and performance.

These observations validate our first hypothesis (see §1), but invalidate the second.

Our findings corroborate those of Garcia et al. (2021b), who showed that vector space models, including BERT, cannot capture the semantic overlap between idiomatic NCs and one or none of their components. Furthering their research, we additionally considered weak idioms, which have a semantic overlap with all of their components, as well as a broader range of idioms, not only NCs.

Our analysis of the effects of POS and the length of idioms suggest that these factors may at least partially explain the model's heightened proficiency at predicting tokens within idioms compared to tokens corresponding to simple lexemes. Nonetheless, this does not explain why CamemBERT is not sensitive to varying levels of idiomaticity among idioms. The very notion of idiomaticity is ambiguous, and the distinction between various types of idiomaticity is often overlooked and tends to be conflated into semantic aspects, i.e., non-compositionality. In our study, we explored both lexical and semantic idiomaticity. Lexical idiomaticity implies that idiomatic tokens exhibit stronger constraints on lexical selection compared to free tokens, i.e., they cannot be replaced by their synonyms while preserving their idiomatic meaning and grammatical correctness. On the other hand, the varying degrees of idiomaticity are indicative of their semantic idiomaticity, which denotes the contribution of internal components to their overall semantic meaning. So CamemBERT's performance in our experiment suggests that in fact the model is more sensitive to lexical idiomaticity than semantic idiomaticity.

This raises questions about other aspects of idiomaticity. Indeed, idioms exhibit idiomaticity on multiple levels simultaneously: lexical, semantic, syntactic, morphological, etc. For instance, *faire la tête* ('sulk', lit. 'make the head') is a strong idiom in French that exhibits not only lexical and semantic idiomaticity, but also prohibits syntactic operations like passivisation, dislocation, etc., as well as morphological inflection to tokens other than the head *faire*. While there is no theoretical consensus on the classification of idiomaticity, our experience may offer valuable insights to address the matter.

In future research, we would like to refine our experiment, extend it to other types of MWEs and explore other forms of idiomaticity. Moreover, we intend to carry out further analyses on language model representations of idiomaticity, exploring additional potential influencing factors such as idiom frequency, or extending our investigation to more complex tasks. We also aim to replicate our experiments with different language models and available datasets in other languages.

6. Acknowledgements

We express our gratitude to the anonymous reviewers for their valuable and constructive feedback. We would like to thank ATILF for the original dataset and our colleagues at OLST for engaging in helpful discussions. Li Liu acknowledges the financial support of the China Scholarship Council (#202008310177).

7. Bibliographical References

Juri Apresjan. 2000. *Systematic Lexicography*. Oxford University Press, Oxford.

ATILF. 2023. Réseau lexical du français (rl-fr). ORTOLANG (Open Resources and TOols for LANGuage)—www.ortolang.fr.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of natural language processing*, 2nd edition, pages 267–292. CRC Press, Boca Raton, FL, USA.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, volume 2, pages 1934–1940, Las Palmas, Spain. ELRA.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL'19: Human Language Technologies*, volume 1, pages 4171–4186, Minneapolis, MN, USA. ACL.

Samin Fakharian and Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 23–32, Online. ACL.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, volume 1, pages 2730–2741, Online. ACL.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. ACL.

Milton King and Paul Cook. 2018. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 345–350, Melbourne, Australia. ACL.

Veronika Lux-Pogodalla. 2014. Integrating lexicographic examples in a lexical network (intégration relationnelle des exemples lexicographiques dans un réseau lexical) [in French]. In *Proceedings of TALN 2014*, volume 2, pages 586–591, Marseille, France. ATALA.

Veronika Lux-Pogodalla and Alain Polguère. 2011. Construction of a French Lexical Network: Methodological Issues. In *First International Workshop on Lexical Resources (WoLeR 2011)*, pages 54–61, Ljubljana, Slovenia.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7203–7219, Online. ACL.

Igor A. Mel'čuk, André P Clas, and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot.

Igor A. Mel'čuk and Alain Polguère. 1987. A formal lexicon in meaning-text theory (or how to do lexica with words). *Computational linguistics*, 13:261–275.

Igor A. Mel'čuk. 2006. Parties du discours et locutions. *Bulletin de la Société de Linguistique de Paris*, 101:29–65.

Igor A. Mel'čuk. 2023. *General phraseology: Theory and practice*. John Benjamins.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

Navnita Nandakumar, Bahar Salehi, and Timothy Baldwin. 2018. A comparative study of embedding models in predicting the compositionality of multiword expressions. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA 2018)*, pages 71–76, Dunedin, New Zealand. ALTA.

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding BERT's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62, Online. ACL.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 3333–3345, Barcelona, Spain (Online). ICCL.

Alain Polguère. 2009. Lexical systems: graph models of natural language lexicons. *Language Resources and Evaluation*, 43:41–55.

Alain Polguère. 2014. From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, 27(4):396–418.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX)*, pages 107–118, online. ACL.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference (CICLing)*, pages 1–15, Mexico. Springer.

Mahtab Sarlak, Yalda Yarandi, and Mehrnoush Shamsfard. 2023. Predicting compositionality of verbal multiword expressions in Persian. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 14–23, Dubrovnik, Croatia. ACL.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Online. INCOMA.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Ye Tian, Isobel James, and Hye Son. 2023. How are idioms processed inside transformer language models? In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 174–179, Toronto, Canada. ACL.

Identification and Annotation of Body Part Multiword Expressions in Old Egyptian

Roberto Antonio Díaz Hernández
University of Jaén
Campus Las Lagunillas, Building D2, Office 352, 23071, Jaén
radiaz@ujaen.es

Abstract

This paper presents the preliminary results of an ongoing study on the diachronic and synchronic use of multiword expressions (MWEs) in Egyptian, begun when I joined the COST Action *Universality, Diversity and Idiosyncrasy in Language Technology* (UniDive, CA21167). It analyzes, as a case study, Old Egyptian body part MWEs based on lexicographic and textual resources, and its aim is both to open up a research line in Egyptology, where the study of MWEs has been neglected, and to contribute to Natural Language Processing studies by determining the rules governing the morpho-syntactic formation of Old Egyptian body part MWEs in order to facilitate the identification of other types of MWEs.

Keywords: Old Egyptian, Multiword Expression, Body Part

1. Introduction

Egyptian is one of the longest lived languages in history. This Afroasiatic language knew the following phases:

- Old Egyptian (ca. 2700–2000 BC).
- Middle Egyptian (ca. 2000–1400 BC).
- Late Egyptian (ca. 1300–700 BC).
- Demotic (7th century BC to 5th century CE).
- Coptic (4th century to 14th century CE).

This paper shows the existence of MWEs in one of the oldest known languages in human history, as they are attested in texts dating from the early third millennium BC (see example 15, below).[1] It focuses on the use of body part MWEs in Old Egyptian, analyzes their typology and identifies rules for their formation. This paper has seven parts. It begins with a brief introduction to the topic (§ 1) and a definition of "body part multiword expression" (§ 2). The methodological approach applied to the identification and annotation of Old Egyptian body part MWEs (§ 3) is followed by examples of each body part noun used in Old Egyptian MWEs (§ 4). A typology of Old Egyptian body part MWEs (§ 5) and an explanation of the rules governing their formation (§ 6) are developed on the basis of the examples collected during the research. Finally, the next phases of this research are outlined in the conclusion (§ 7).

2. A Definition of a Body Part Multiword Expression

It is assumed that "body" and "body parts" are universal concepts (Wierzbicka, 2007) which can be used with a metonymic and metaphoric meaning (Ganfi, Piunno and Mereu, 2023). A body part MWE may be defined as a sequence of at least two lexicalized components, one of which is a body part name, whose semantic idiosyncrasy results from the association between the body part with a figurative meaning and another component(s) (*cf.* the definition of MWE in Savary *et al.*, 2018; Baldwin and Kim, 2010). Body part MWEs are common in modern and ancient languages, e.g.:

1. English:

LM: "Listen to your heart."[2]
FT: "Act according to your feelings."

2. Latin (Plaut., *Asin.* 729):

nec	caput	nec	pes
neither-neg	head	nor-neg	foot

LM: "Neither head nor foot."
FT: "Completely wrong."

3. Arabic:

القلب	ضعيف
al-qalb	*ḍʿīf*
the heart-M.SG.DET	weak-M.SG

LM: "A weak one of heart."
FT: "A coward."

3. Methodology

Although multiple forms of figurative language, such as simile and metaphor have been extensively studied in Egyptology,[3] the study of MWEs remains unexplored. Old Egyptian body part MWEs was chosen as a case study for this work because of the occasional metonymic and metaphoric use of body part nouns (see § 2, above)—a factor that facilitates the identification of MWEs in any language (see examples 1–3, above). Lexical compounds with an idiosyncratic meaning consisting of a body part noun in a close relationship with its head word were considered as MWEs, as for example:

[1] Earlier instances of MWEs may be found in Sumerian texts from the Early Dynastic Period (ca. 2900 BC).

[2] LM stands for "literal meaning" and FL for "free translation".

[3] For the state of the art in Egyptian figurative language, see Hsu 2023.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 24–32. May 25, 2024.

4. CG 20543, 5:

ꜥḳ	*ỉb*	*nb.t*	*=f*
enter:PTCP(M.SG)	heart-M.SG	mistress-F.SG	=3SG.M

LT: “One who enters the heart of his mistress.”
FT: “A confidant of his mistress.”

Metaphorical expressions used to establish a figurative comparison of two entities by means of a “comparison marker”, such as *mr* “like” in Old Egyptian were disregarded in this research, for example:

5. Pyramid Texts § 293a:

ỉgp	*=k*	*mr*	*ꜥḥꜥ.w*
soar-SBJV	=2SG.M	like:PREP	heron-M.SG

LM: “You shall soar (skyward) as a heron.”
FT: “You shall fly over the clouds.”

A fuzzy boundary represents the case where the body part noun has a metonymic meaning, while the head word retains its literal meaning. Such cases were included as MWEs in this research (see identification tests 2 and 3, below), as for example:

6*. Pyramid Texts § 1592e:

mrr	*ỉb*	*=f*	*ỉm*	*(=f)*
love:REL.PRS	heart-M.SG	=3SG.M	in:PREP	(=3SG.M)

LM: “(... any place) which his heart (i.e. will) loves.”
FT: “(... any place) which he desires.”

Body part MWEs are clearly identified when its figurative meaning results from the close association of the body part noun with its head word, as for example:

7. Pyramid Texts § 22b:

ḳb	*ỉb*	*=k*
be cool:SBJV	heart-M.SG	=2SG.M

LM: “Your heart may be cool.”
FT: “You may be calm (i.e. satisfied).”

In Egyptian the figurative meaning of a body part MWE is often related to the idiosyncrasy of this language, as the following example shows:

8. Pyramid Texts § 417b:

ỉm(.ï) rṭ “one who is in the foot”

The figurative meaning of this expression is “enemy”, for it derives from the Egyptian custom of decorating sandals with the image of foes:

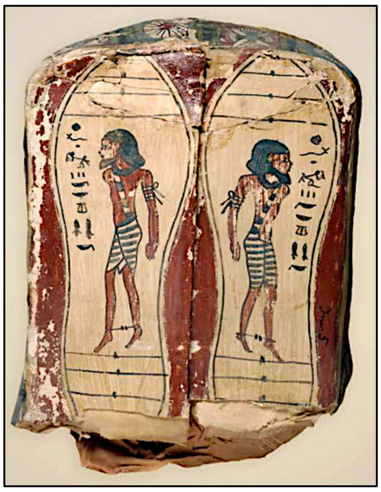
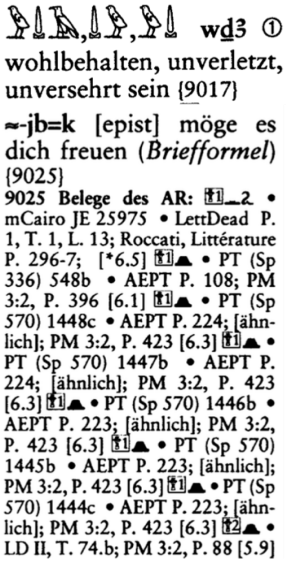


Fig. 1: Foot-end of mummy cartonnage (Veldmeijer, 2014)

Although MWEs are not identified in Hannig’s Old Egyptian dictionary, it provides extensive references to the meaning of each Egyptian word and lexical compound:



wḏꜣ ①
wohlbehalten, unverletzt, unversehrt sein {9017}

≈-jb=k [epist] möge es dich freuen (*Briefformel*) {9025}

9025 Belege des AR: 1_2 • mCairo JE 25975 • LettDead P. 1, T. 1, L. 13; Roccati, Littérature P. 296-7; [*6.5] 1▲ • PT (Sp 336) 548b • AEPT P. 108; PM 3:2, P. 396 [6.1] 1▲ • PT (Sp 570) 1448c • AEPT P. 224; [ähnlich]; PM 3:2, P. 423 [6.3] 1▲ • PT (Sp 570) 1447b • AEPT P. 224; [ähnlich]; PM 3:2, P. 423 [6.3] 1▲ • PT (Sp 570) 1446b • AEPT P. 223; [ähnlich]; PM 3:2, P. 423 [6.3] 1▲ • PT (Sp 570) 1445b • AEPT P. 223; [ähnlich]; PM 3:2, P. 423 [6.3] 1▲ • PT (Sp 570) 1444c • AEPT P. 223; [ähnlich]; PM 3:2, P. 423 [6.3] 2▲ • LD II, T. 74.b; PM 3:2, P. 88 [5.9]

Fig. 2: Textual references to the MWE *wḏꜣ ib* “be happy” (Hannig 2003: 398-399)

I checked the references of body part nouns potentially used in MWEs against the editions of hieroglyphic texts. Instances of body part nouns with a literal meaning have been disregarded (see validation test 1, below), while instances of body part nouns in figurative association with other words have been considered body part MWEs according to the definition given in section 2 (see above). In addition, I used the textual database of the *Thesaurus Linguae Aegyptiae* to find further instances of body part MWEs in Old Egyptian texts. After selecting and entering them into an Excel list, I manually annotated the most eloquent examples of Old Egyptian body part MWEs in a Word file for lack of a digital resource.[4] Such examples have a clear meaning and syntactic structure. As it can be seen here, they were annotated with the reference source and following the Leipzig Glossing Rules.[5] They are provided with a literal meaning (LM) and a free translation (FT), for example:

9. Pyramid Texts § 293a:

nčm	*ỉb*	*n(.ỉ)*	*[Wnỉś]*
be sweet-SBJV	heart-M.SG	of-M.SG	Unas-KN

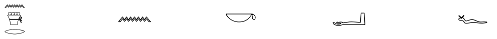
LM: “The heart of [Unas] shall be sweet.”
FT: “[Unas] shall be kind.”

The selection and identification of Old Egyptian body part MWEs was carried out using a series of verification tests:

Test 1. Does the body part noun have a literal meaning?

— Yes ⇒ It is not an MWE, for example:

10. Pyramid Texts § 49 Nt:



nčr	*n*	*=k*	*ʿ*	*=f*
seize:IMP	for:PREP	=2SG.M	arm-M.SG	=3SG.M

LM: “Seize for yourself his arm.”
FT: “Seize his arm!”

— No ⇒ Test 2.

Test 2. Does the body part noun have a metonymic meaning?

— Yes ⇒ It is a potential MWE ⇒ Test 3. Ex.:

11. Pyramid Texts § 1675b:

śšm	*čw*	*ỉb*	*=k*
guide:SBJV	=2SG.M	heart-M.SG	=2SG.M

LT: “Your heart shall guide you.”
FT: “Your will shall guide you.”

— No ⇒ It is not an MWE, see test 1.

Test 3. Is the body part noun used with an idiosyncratic meaning in close syntactic relationship with a head word?

— Yes ⇒ It is an MWE, for example:

12. Pyramid Texts § 116a:

ỉnč	*(=ỉ)*	*ḥr*	*=k*
ask:SBJV	=1SG	face-M.SG	=2SG.M

LT: “May (I) ask your face.”
FT: “Hail to you!”

— No. It is a fuzzy MWE consisting of a body part noun with a metonymic meaning (see test 2). It has been marked with an asterisk in this paper, see ex. 6*, 16*, 27*, 32* and 43*.

Test 4. Is the body part noun used in a lexicalized expression with an idiosyncratic meaning?

— Yes ⇒ It is an MWE. This is the usual case for complex prepositions (CPs), for example:

13. Pyramid Texts § 54b:

fꜣ	*ḫft*	*ḥr*	*=f*
lift up:IMP	in front of:PREP	face-M.SG	=3SG.M

LM: “Lift up in front of his face.”
FT: “Lift up before him”

It should also be noted that body part MWEs are occasionally attested in some scenes, which are annotated here in order to illustrate their meaning, for example:

14. Davies, 1900, pl. III, *cf.* fig. 3:

wčꜣ	*ỉb*	*=k*	*(ỉ)r*	*šy*
be hale:SBJV	heart-M.SG	=2SG.M	concerning:PREP	crocodile-M.SG

LM: “Your heart shall be hale concerning the crocodile.”
FT: “You shall be happy of having escaped from the crocodile.”

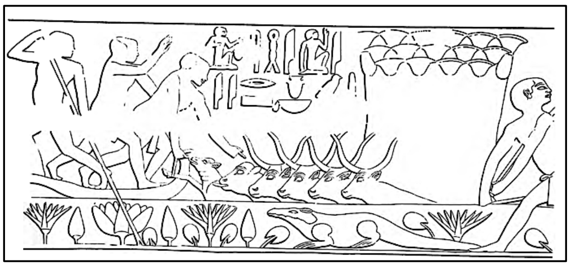


Fig. 3: A cow escapes from the crocodile

In Old Egyptian the frequency of body part MWEs varies depending on the body part noun—the commonest body part MWEs are those consisting of *ỉb* “heart” (no less than 63 types of MWEs)[6] and ʿ “arm” (no less than 24 types of MWEs), while the less common body part MWEs are those consisting of *ỉr.t* “eye”, *ʿn.t* “nail” and *ḫpš* “biceps” which are attested in less than five types of MWEs. The following Old Egyptian body part nouns are used in MWEs (see examples in § 4, below):

[4] I am working on the first treebank of Egyptian sentences syntactically analyzed in Universal Dependencies. Its initial release is planned for Mai 2024. This treebank will include MWEs to be published as a corpus in PARSEME.

[5] Di Biase-Dyson, Kammerzell and Werning (2009) adapted the Leipzig Glossing Rules for the study of Egyptian texts.

[6] I discussed the use of Old Egyptian MWEs containing *ỉb* in a poster I presented at the second general meeting of UniDive (Università di Napoli “L’Orientale”, 8–9 February 2024). The 63 types of Old Egyptian MWEs containing *ỉb* are analyzed one by one in my forthcoming article “Old Egyptian Multiword Expressions consisting of a head word and *ỉb* ‘heart’”.

Spelling	Transcription	Literal meaning
	i͗wf	flesh
	i͗b	heart
	i͗r.t	eye
	ʿ	arm
	ʿn.t	nail
	r᾽	mouth
	rmn	shoulder
	rṭ	foot
	ḥr	face
	ḥꜣ.t	forehead
	ḥꜣ.ti͗	heart
	ḫpš	strong arm (biceps)
	ẖ.t	belly
	sꜣ	back
	šni͗	hair
	tp	head
	č.t	body
	čbʿ	finger
	čr.t	hand

Table 1: Body part nouns used in Old Egyptian MWEs

4. Evidence

The earliest instances of body part MWEs in Egyptian date from the Early Dynastic Period (ca. 2900–2730 BC), for example:

15. Petrie, 1901 (vol. 2, pl. III), *cf.* fig. 4:

i͗mꜣ	*i͗b*
be kind:PTCP (M.SG)	heart-M.SG

LM: "One who is kind of heart."
FT: "A well-liked one."

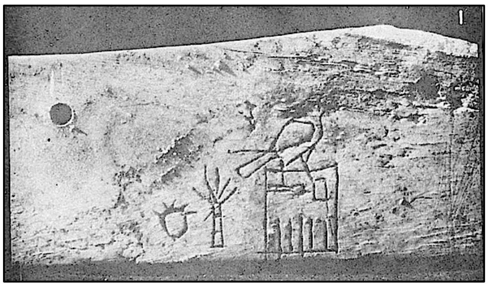


Fig. 4: An Abydos tablet

As the following examples show, body part nouns listed in table 1 (see above) are used to form Old Egyptian MWEs.

16. Example of *i͗wf* (Sethe, 1933, 14,5):

prr	*=f*	*r*	*=f*	*m*	*i͗(w)f*	*=f*
go:PRS	=3SG.M	PTCL	=3SG.M	with:PREP	flesh-M.SG	=3SG.M

LM: "He goes off for him with his flesh."
FT: "He goes off, certainly at his own risk."

17. Example of *i͗b* (Brunner, 1937, 62,79):

ẖꜣk	*i͗b*	*nb*
be hostile-PTCP(M.SG)	heart-M.SG	every-M.SG

LM: "Everyone who is hostile of heart."
FT: "Any evil-minded person."

18. Example of *i͗r.t* (Černý, 1961, 7):

n	*psg*	*(=i͗)*	*m*	*i͗r.ti͗*	*n(.t)*	*nfr*
not:NEG	spit:PST	=1SG	in:PREP	eye-F.DU	of-F	good-M.SG

LM: "(I) did not spit in the two eyes of a good one."
FT: "(I) did not spit on the eyes of a good man (i.e. I did not humiliate a good man)."

19. Example of *ʿ* (Pyramid Texts, § 213a):

m	*ẖnw*	*ʿ*	*=k*
in:PREP	interior-M.SG	arm-M.SG	=2SG.M

LM: "(...) in the interior of your arm."
FT: "(...) within your embrace."

20. Example of *ʿn.t* (Moussa/Altenmüller, 1977, 79 and fig. 10), *cf.* fig. 5:

i͗r[.t]	*ʿn.(w)t*
make:INF	nail-F.PL

LM: "Making the nails."
FT: "Cutting nails (or manicure/pedicure)."

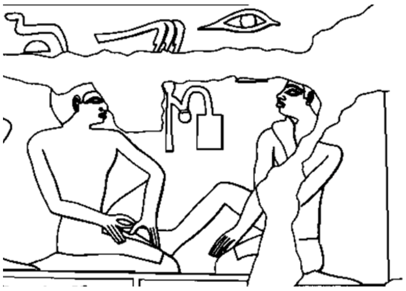


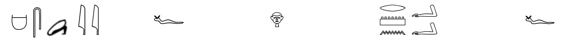
Fig. 5: A man pedicuring another man

21. Example of *r᾽* (Pyramid Texts, § 1299a):

rč	*=k*	*r᾽*	*=k*	*n(.i͗)*	*Rʿ*
give:FUT	=2SG.M	mouth-M.SG	=2SG.M	of-M.SG	Ra-GN

LM: "You will give your mouth to Ra."
FT: "You will speak with Ra."

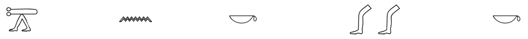
22. Example of *rmn* (Pyramid Texts, § 813a):



ḥmś.y	*⸗f*	*ḥr*	*rmn(.wỉ)*	*⸗f*
sit:FUT	=3SG.M	on:PREP	shoulder-M.DU	=3SG.M

LM: “He will sit on his two shoulders.”
FT: “He will sit himself beside him.”

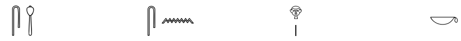
23. Example of *rṭ* (Kanawati 1997, fig. 41):



ỉč	*n*	*⸗k*	*rṭ(.wỉ)*	*⸗k*
take:IMP	for:PREP	=2SG.M	foot-M.DU	=2SG.M

LM: “Take for you your two feet.”
FT: “Move!”

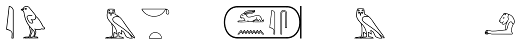
24. Example of *ḥr* (Pyramid Texts, § 613a):



ś:ḥč	*⸗śn*	*ḥr*	*⸗k*
make bright:SUBJV	=3PL	face-M.SG	=2SG.M

LM: “They shall make your face bright.”
FT: “They shall make you glad.”

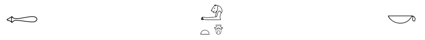
25. Example of *ḥꜣ.t* (Pyramid Texts, § 407d):



ỉw	*mk.t*	*Wnỉś*	*m*	*ḥꜣ.t*
PTCL	place-F.SG	Unas-KG	at:PREP	forehead-F.SG

LM: “The place of Unas is at the forehead.”
FT: “Unas’ place is ahead.”

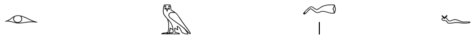
26. Example of *ḥꜣ.tỉ* (Pyramid Texts, § 2024a):



ꜥꜣ	*ḥꜣ.tỉ*	*⸗k*
be great:SUBJV	heart-M.SG	=2SG.M

LM: “Your heart shall be great.”
FT: “Be proud!”

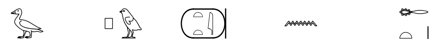
27*. Example of *ḫpš* (Fischer, 1961, 47):



ỉr	*m*	*ḫpš*	*⸗f*
act-PTCP(M.SG)	with:PREP	biceps-M.SG	=3SG.M

LM: “One who acted with his biceps.”
FT: “One who acts on his own.”

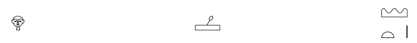
28. Example of *ẖ.t* (Pyramid Texts, § 1c):



sꜣ	*pw*	*Ttỉ*	*n(.ỉ)*	*ẖ.t*	*(⸗ỉ)*
son-M.SG	COP	Teti-KN	of-M.SG	belly-F.SG	=1SG

LM: “Teti is the son of (my) belly.”
FT: “Teti is (my) bodily (i.e. biological) son.”

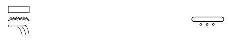
29. Example of *śꜣ* (Sethe, 1933, 111,8):



ḥr	*śꜣ*	*ḫꜣś.t*
on:PREP	back-M.SG	foreign land-F.SG

LM: “(...) on the back of the foreign land.”
FT: “(...) at the far end of the foreign land.”

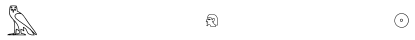
30. Example of *šnỉ* (Petrie, 1900, pl. XXVB):



šnỉ	*tꜣ*
hair-M.SG	earth-M.SG

LM: “Hair of the earth.”
FT: “Vegetation.”

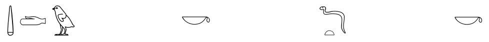
31. Example of *ṭp* (Pyramid Texts, § 989a):



m	*ṭp*	*hrw*
in:PREP	head-M.SG	day-M.SG

LM: “(...) in the head of the day.”
FT: “(...) at dawn.”

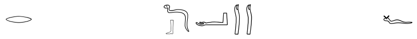
32*. Example of *č̣.t* (Pyramid Texts, § 762b):



mṭw	*⸗k*	*č̣.t*	*⸗k*
speak:SUBJV	=2SG.M	body-M.SG	=2SG.M

LM: “You shall speak (of) your body.”
FT: “You shall speak (of) yourself.”

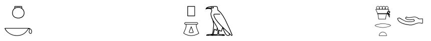
33. Example of *č̣bꜥ* (Pyramid Texts, § 372a):



r	*č̣bꜥ(.wỉ)*	*⸗f*
to:PREP	finger-M.DU	=3SG.M

LM: “(...) to his two fingers.”
FT: “(...) at his side.”

34. Example of *č̣r.t* (Brunner, 1937, 42,3):



ỉnk	*pgꜣ*	*č̣r.t*
1SG	open-PTCP(M.SG)	hand-F.SG

LM: “I am one who opens the hand.”
FT: “I am a generous one.

5. Typology

Old Egyptian body part MWEs can be classified according to universal typology as nominal, prepositional and verbal.[7] In nominal multiword expressions (NMWEs) the head word accompanying the body part noun can be a noun, an infinitive, an adjective or a participle. The head word of prepositional multiword expressions (PMWEs) can only be a preposition. In verbal multiword expressions (VMWEs) the head word must be a verb form (except if it is a nominalized verb form which is considered an NMWE).

5.1 Nominal Multiword Expressions

A body part noun can be the head or the modifier of an NMWE. If it is the former, it usually means a physical object, for example:

[7] See Baldwin and Kim, 2010, 274–279.

35. Goedicke, 1994, 73, I.9, *cf.* fig. 6:

ʿ	*m*	*ḫt*
arm-M.SG	of:PREP	wood-M.SG

LM: “An arm (made) of wood.”
FT: “An incense burner (in the shape of an arm).”

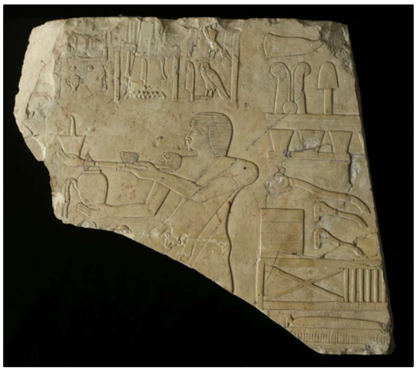


Fig. 6: A ritualist holding an incense burner (Walters Art Museum 22216)

If the body part noun is used as a modifier, the head of the NMWE can be a noun, an infinitive, and an adjective or a participle:

36. Example of a noun as the head of an NMWE (Junker, 1943, fig. 43):

čꜣ(.w)	*śrf*	*i͗b*
man-M.SG	warm-M.SG	heart-M.SG

LM: “A warm man of heart.”
FT: “A hard-working man.”

37. Example of an infinitive as the head of an NMWE (Paget, 1898, pl. XXXVIII), *cf.* fig. 7:

i͗n.t	*rṭ*
bring:INF	foot-M.SG

LM: “Bringing the foot.”
FT: “Erasing the footprint (a ritual ceremony).”

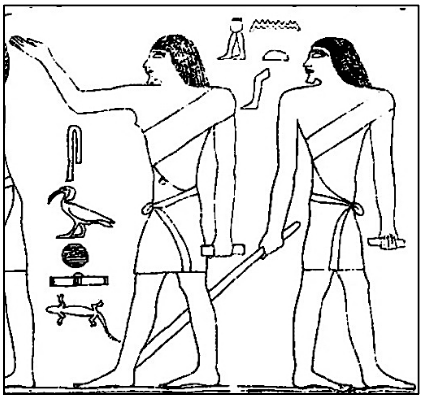


Fig. 7: A ritualist “erasing the footprint”

38. Example of an adjective as the head of an NMWE (Pyramid Texts, § 195c):

nfr	*w(i͗)*	*ḥr*	*=č*
beautiful-M.SG	PTCL	face-M.SG	=2SG.F

LM: “How beautiful is your (f.) face.”
FT: “How nice is to see you.”

39. Example of a participle as the head of an NMWE (Pyramid Texts, § 1a):

Tti͗	*wp*	*ẖ.t*	*(=i͗)*
Teti-KN	open-PTCP(M.SG)	belly-F.SG	=1SG

LM: “(...) Teti who opened (my) belly.”
FT: “(...) Teti, (my) first-born.”

5.2 Prepositional Multiword Expressions

Body part nouns are used as modifiers in prepositional multiword expressions. Two types of PMWEs can be found in Old Egyptian: prepositional idioms (PIs) and complex prepositions (CPs).

40. Example of a prepositional idiom (Sethe, 1933, 162,11):

ẖr	*ʿ*	*śꜣ*	*(=i͗)*	*śmś.w*
under:PREP	arm-M.SG	son-M.SG	=1SG	eldest-M.SG

LM: “(...) under the arm of (my) eldest son.”
FT: “(...) under the care of (my) eldest son.”

41. Example of a complex preposition (Sethe, 1933, 126,2):

m	*śꜣ*	*=f*
in:PREP	back-M.SG	=3SG.M

LM: “(...) in his back.”
FT: “(...) behind him.”

5.3 Verbal Multiword Expressions

Body part nouns are also modifiers in VMWEs. Old Egyptian body part VMWEs are usually verbal idioms (IDs) consisting of a verb as a head and a body part noun with a figurative meaning, for example:

42. Pyramid Texts, § 425a:

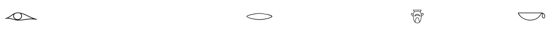
mḥ.n	*=f*	*rʾ*	*n(.i͗)*	*Wni͗ś*
fill:PST	=3SG.M	mouth-M.SG	of-M.SG	Unas-KN

LM: “(...) he filled the mouth of Unas.”
FT: “(...) he fed Unas.”

Light Verb Constructions consisting of a “light” verb and a noun denoting an event or a state, such as “make a speech”[8] are hardly found in Old Egyptian body part VMWEs. However, the metonymic meaning of body part nouns occasionally refers to an action which modifies the meaning of the expression, for example:

[8] *Cf.* Savary *et al.*, 2018, 99 and 102; Baldwin and Kim, 2010, 277.

43*. Duell, 1938, pl. 162, *cf.* fig. 8:



ỉr	*(=ỉ)*	*r*	*ỉb*	*=k*
do:SBJV	=1SG	according to-PREP	heart-M.SG	=2SG.M

LM: "(I) shall do according to your will (lit.: heart)."
FT: "(I) will do what you want."

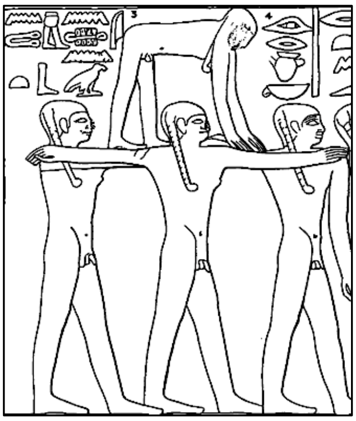


Fig. 8: A boy following the instructions of his friends

6. Formation Rules

The formation of Old Egyptian body part MWEs follows strict morpho-syntactic rules, which are useful not only for understanding how an MWE was used in Old Egyptian, but also for identifying other types of MWEs. Five formation rules are derived from the morpho-syntactic analysis of Old Egyptian body part MWEs:

1) A verb stem in a VMWE can be transformed into an infinitive in an NMWE, *cf.*:

44. Example of a VMWE consisting of the subjunctive *ꜣw + ỉb* (Pyramid Texts, § 715c):

ꜣw	*ỉb*	*n(.ỉ)*	*nčr(.w)*	*m*	*Ttỉ*
be long:SBJV	heart-M.SG	of-M.SG	god-M.PL	in-PREP	Teti-KN

LM: "The heart of the gods shall be long in Teti."
FT: "The gods shall be glad over Teti."

45. Example of an NMWE consisting of the infinitive *ꜣw.t + ỉb* (Pyramid Texts § 1175a):

tꜣ	*m*	*ꜣw.t*	*ỉb*
earth-M.SG	in:PREP	length-F.SG	heart-M.SG

LM: "The earth is in length of heart."
FT: "The earth is in joy."

2) A verb stem in a VMWE can be transformed into a participle in an NMWE, *cf.*:

46. Example of a VMWE consisting of the verb form *ỉ:wn* + *ḥr* (Pyramid Texts, 391c):

ỉ:wn	*ḥr*	*nčr*	*n*	*Wnỉś*
open:PASS.FUT	face-M.SG	god-M.SG	to:PREP	Unas-KN

LM: "The face of the god will be open to Unas."
FT: "The god will view the king with favour."

47. Example of an NMWE consisting of the participle *wn + ḥr* (Sethe, 1933, 149,1):

wn	*ḥr*	*n*	*{ḫ}<č̣>ꜣm(.w)*
open-PTCP(M.SG)	face-F.SG	to:PREP	troops-M.PL

LM: "One who opens the face to the troops."
FT: "One who views the troops with favour."

Note that deverbal constructions resulting from a VMWE into an NMWE are also found in other languages, such as English:

"She makes decisions quickly" > "She is a quick decision maker" (see Savary *et al.*, forthcoming).

3) A preposition in a PMWE can be transformed into a nisba adjective in an NMWE,[9] *cf.*:

48. Example of a PMWE consisting of the preposition *ẖr + ʿ* (Sethe, 1933, 162,11):

ẖr	*ʿ*	*śꜣ*	*(=ỉ)*	*śmś.w*
under:PREP	arm-M.SG	son-M.SG	=1SG	eldest-M.SG

LM: "(...) under the arm of (my) eldest son."
FT: "(...) under the care of (my) eldest son."

49. Example of an NMWE consisting of the nisba adjective *ẖr.(ỉ)w + ʿ* (Pyramid Texts, § 1236b):

ẖr.(ỉ)w	*ʿ*	*Wśr(.w)*
those who is under-M.PL	arm-M.SG	Osiris-GN

LM: "Those who are under the arm of Osiris."
FT: "Those who are under the care of Osiris."

4) The nisba adjective resulting from a preposition can be used as a noun in an NMWE, for example:

50. Goedicke, 1968, 27:

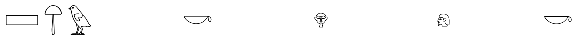
ẖr.(ỉ)	*ʿ*
one who is under-M.SG	arm-M.SG

LM: "One who is under the arm."
FT: "One who is under the care (i.e. assistant)."

Note that the usual transformation of a preposition in a PMWE into a nisba adjective or a noun in an NMWE is an idiosyncratic feature of Old Egyptian hardly found in other languages. This is a common way for the formation of Egyptian titles, for example the title *ḥr(.ỉ) ṯp* "great chief" is derived from the PMWE *ḥr ṯp* "on the head", *cf.*:

[9] In Semitic languages, such as Arabic, "nisba" is used to label an ending added to nouns, and rarely to prepositions and pronouns, to form (relative) adjectives and nouns (see Schulz 2010, 86). The addition of the nisba ending to prepositions to form adjectives and nouns is a common feature in Egyptian.

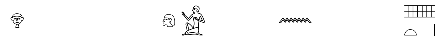
51. Pyramid Texts 1487a:



šw	*⸗k*	*ḥr*	*tp*	*⸗k*
shade-M.SG	=2SG.M	on:PREP	head-M.SG	=2SG.M

LM: “Your shade is on your head.”
FT: “Your shade is over you.”

52. Sethe 1933, 254,4:

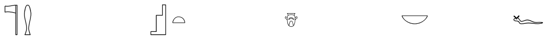


ḥr(.ỉ)	*tp*	*n(.ỉ)*	*śpꜣ.t*
one who is on-M.SG	head-M.SG	of-M.SG	nome-F.SG

LM: “One who is on the head of the nome.”
FT: “Great chief of the nome (i.e. nomarch).”

5) An NMWE consisting of a noun as its head word can be transformed into a PMWE by adding a preposition before the noun, *cf.*:

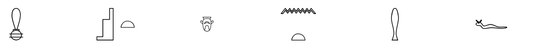
53. Example of an NMWE consisting of the nouns *ś.t* + *ỉb* (CG 1485):



ḥm-nčr	*ś.t*	*ỉb*	*nb*	*⸗f*
priest-TITLE	place-F.SG	heart-M.SG	lord-M.SG	=3SG.M

LM: “The priest of the place of the heart of his lord.”
FT: “The priest beloved of his lord (i.e. the favourite priest of his lord).”

54. Example of a PMWE consisting of the preposition *mr* + *ś.t ỉb* (Sethe, 1933, 56,19):



mr	*ś.t*	*ỉb*	*n.t*	*ḥm*	*⸗f*
like:PREP	place-M.SG	heart-M.SG	of-F.SG	majesty-M.SG	=3SG.M

LM: “(I used to act) like the place of the heart of his majesty.”
FT: “(I used to act) at the request of his majesty.”

7. Conclusion

This research leads to the following preliminary results:

1) The existence of MWEs is indisputable in Old Egyptian, which means that they are as old as the Pyramids of Giza.

2) Body part nouns are used in Old Egyptian to form MWEs, which means that Old Egyptian phrases containing a body part noun with a metonymic meaning are potential candidates to be identified as MWEs.

3) The typology of body part MWEs in Old Egyptian is similar to that applying to MWEs in other languages.

Research on MWEs in Egyptian will be continued in these two phases:

1) Publication of the selected examples in PARSEME after having annotated them manually in the Universal Dependencies treebank "Egyptian-UJaen".

2) Identification and classification of new Old Egyptian MWEs following the rules discussed in this paper and the identification tests suggested in Savary *et al.*, 2018.

Once the synchronic study of MWEs in Old Egyptian is completed, their analysis in later stages of Egyptian will follow in order to detect changes during their historical development. This will contribute not only to the confirmation of the universal categorization of MWEs, based mostly on modern Indo-European languages, but also to the development and refinement of universal rules concerning the formation of MWEs. The end result of this research will be a manually annotated digital corpus of Egyptian MWEs published in PARSEME and a lexicon of Egyptian MWEs.

8. Acknowledgments

I thank UniDive for having introduced me to the study of NLP.

9. Bibliographical References

Baldwin, T. and Kim, S. N. (2010). Multiword Expressions. In: N. Indurkhya and F. J. Damerau (eds.) *Handbook of Natural Language Processing*. Boca Raton, London, New York, pp. 267–292.

Brunner, H. (1937). *Die Texte aus den Gräbern der Herakleopolitenzeit von Siut*. Glückstadt.

CG 1485 = Borchardt, L. 1937–1964. *Catalogue général des antiquités égyptiennes du Musée du Caire. Denkmäler des Alten Reiches*. Cairo.

CG 20543 = Lange, H. O. and Schäfer, H. 1902–1925. *Catalogue général des antiquités égyptiennes du Musée du Caire. Grab- und Denksteine des Mittleren Reichs* (4 vols.), Berlin.

Černý, J. (1961). The Stela of Merer in Cracow. *The Journal of Egyptian Archaeology* 47:5–9.

Davies, N. G. (1900). *The Mastaba of Ptahhetep and Akhethetep at Saqqarah*. London.

Di Biase-Dyson, C., Kammerzell, F. and Werning, D. A. (2009). Glossing Ancient Egyptian. Suggestions for adapting the Leipzig Glossing Rules. *Lingua Aegyptia* 17:343–366.

Duell, P. (1938). *The Mastaba of Mereruka. Part II*. Chicago.

Fischer, H. G. (1961). The Nubian Mercenaries of Gebelein during the First Intermediate Period. *Kush. Journal of the Sudan Antiquities Service* 9:44–80.

Ganfi, V., Piunno, V. and Mereu, L. (2023). Body part metaphors in phraseological expressions. *Languages in Contrast* 23(1):1–33.

Goedicke, H. (1968). Four Hieratic Ostraca of the Old Kingdom. *The Journal of Egyptian Archaeology* 54:22–30.

Goedicke, H. (1994). A Cult Inventory of the Eight Dynasty from Coptos. *Mitteilungen des Deutschen Archäologischen Instituts, Abteilung Kairo* 50:71–85.

Hsu, S.-W. (2023). Figurative Language. In: A. Stauder, W. Wendrich (eds.) *UCLA Encyclopedia of Egyptology*, Los Angeles.

Junker, H. (1943). *Gîza VI. Die Maṣṭabas des Nfr (Nefer), Ḳdfjj (Kedfi), Kꜣḥjf (Kaḥjef) und die westlich anschließenden Grabanlagen,* Vienna and Leipzig.

Kanawati, N. and Hassan, A. (1997). *The Teti Cemetery at Saqqara. Volume II. The Tomb of*

Ankhmahor. The Australian Centre for Egyptology Reports 9, Wiltshire.

Moussa, A. and Altenmüller, H. (1977). *Das Grab des Nianchchnum und Chnumhotep*. Mainz am Rhein.

Paget, R. F. 1898. The Tomb of Ptah-hetep. In: J. E. Quibell (ed.) *The Ramesseum*. London.

Petrie, W. M. F. (1900). *Dendereh. Extra Plates*. London.

Petrie, W. M. F. (1901). *The Royal Tombs of the Earliest Dynasties. Part II*. London.

Pyramid Texts = Sethe, K. (1908–1922). *Die altägyptischen Pyramidentexte nach den Papierabdrücken und Photographien des Berliner Museums* (4 vols.). Leipzig.

Savary, A. *et al.* (2018). PARSEME multilingual corpus of verbal multiword expressions. In: S. Markantonatou *et al*. (eds.) *Multiword Expressions at Length and in Depth. Extended Papers from the MWE 2017 Workshop*. Berlin, pp. 87–148.

Savary, A. *et al.* forthcoming. Guidelines for Nominal MWEs.

Schulz, E. (2010). *A Student Grammar of Modern Standard Arabic*, Cambridge.

Sethe, K. (1933). *Urkunden des Alten Reichs*. Leipzig.

Veldmeijer, A. J. (2014). *Footwear in Ancient Egypt: The Medelhavsmuseet Collection*. Världskulturmus eerna.

Wierzbicka, A. (2007). Bodies and their parts: An NSM approach to semantic typology. *Language Sciencies*, 29:14–65.

10. Language Resource References

Hannig, R. 2003. *Ägyptisches Wörterbuch I. Altes Reich und Erste Zwischenzeit*. Mainz am Rhein.

Thesaurus Linguae Aegyptiae: https://thesaurus-linguae-aegyptiae.de/home

Fitting Fixed Expressions into the UD Mould: Swedish as a Use Case

Lars Ahrenberg
Department of Computer and Information Science
Linköping University
lars.ahrenberg@liu.se

Abstract

Fixed multiword expressions are common in many, if not all, natural languages. In the Universal Dependencies framework, UD, a subset of these expressions are modelled with the dependency relation *fixed*, targeting the most grammaticalized cases of functional multiword items. In this paper we perform a detailed analysis of 439 expressions modelled with *fixed* in two Swedish UD treebanks in order to reduce their numbers and fit the definition of *fixed* better. We identify a large number of dimensions of variation for fixed multiword expressions that can be used for the purpose. We also point out several problematic aspects of the current UD approach to multiword expressions and discuss different alternative solutions for modelling fixed expresions. We suggest that insights from Constructional Grammar (CxG) can help with a more systematic treatment of fixed expressions in UD.

Keywords: Multiword expressions, fixed expressions, constructions, Swedish

1. Introduction

Multiword expressions (MWEs) are ubiquitous in many, if not all, natural languages. They are usually divided into different classes with fixed, word-like expressions at one end and flexible phrase- and clause-like expressions at the other. Common English examples of these two kinds are illustrated in (1) and (2):

(1) *at first, by and large, of course*
(2) *give X the creeps, beat around the bush*

How do you search for MWEs in a treebank annotated in the Universal Dependencies (UD) framework? That would depend on the type of MWE you are interested in. UD offers three relations to represent MWEs: *compound, flat* and *fixed* (de Marneffe et al., 2021). The first is focused on compounding of nouns and other content words, the second on fixed expressions with similar behavior as function words, and the third primarily on multiword names. For definitions see Table 1. If your interest is with the flexible ones, however, you would have to use the key words of the MWE such as *creeps* or *around the bush*, as there is no particular relations devoted to them; they are annotated the same way as compositional phrases and clauses. Alternatively, you can turn to treebanks with more flexible annotations such as those developed in the PARSEME project with special annotations for verbal multiword expressions (Savary et al., 2023a).

The stated purpose of UD is to develop crosslinguistically consistent morphosyntactic annotation for as many languages as possible. The main purposes are to support research in language typology and natural-language processing, parsing in particular. Given that MWEs sometimes show deviant morphosyntactic behaviour and that the knowledge of MWEs crosslinguistically appears to be scarce (Masini, 2019) we can argue that MWEs should be given adequate representations in UD annotation. Then it is a problem that it does not cover all types of MWEs. While this problem has been recognized (Savary et al., 2023b), no solution has been agreed upon so far.

A framework that places MWEs at the center of linguistic modelling is Construction Grammar (CxG) (Fillmore et al., 1988; Booij, 2017; Hoffmann, 2022). The most radical view of CxG holds that everything in language, from morphs to sentences, are instances of form-meaning pairs of the same basic kind, called constructions. A form is a pattern of some sort and the meaning may be more or less specific. In contrast, UD only recognizes the existence of certain MWEs and by using the syntactic level of annotation it actually blurs the fact that MWEs often have a transparent syntactic structure; MWEs don't have to be syntactically deviant.

The empirical basis of the paper is a detailed analysis of the formal and structural variation in MWEs currently annotated as *fixed* in two Swedish UD treebanks. All expressions in this dataset have been annotated for the type of variation they accept, their distribution if regarded as a UD word, and for their structure. The latter aspect takes inspiration from the treatment of MWEs in Construction Grammar, in particular the idea that structures can enter into hierarchical relations. While the data is primarily taken from Swedish they illustrate general types of problems in relation to fixed MWEs. Comparisons are made with the use of *fixed* in UD treebanks for English.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 33–42. May 25, 2024.

Relation	Definition
compound	any kind of word-level compounding (noun compound, serial verb, phrasal verb)
fixed	fixed multiword expression; links elements of grammaticalized expressions that behave as function words or short adverbials
flat	flat multiword expression; links elements of headless semi-fixed multiword expressions like names

Table 1: Definitions of the three dependency relations used for MWEs in UD cited from (de Marneffe et al., 2021)[266]

The paper is structured as follows. The next section provides background on fixed MWEs as found in general overviews, in Usage CxG, and in UD. Section 3 presents our dataset and how it has been annotated. In Section 4 we review a number of common types of fixed MWEs found in the dataset and discuss how they can be analysed with or without the *fixed* relation. Section 5 proposes alternative ways to annotate them in UD. Section 6, finally, holds the conclusions.

2. Multiword Expressions in Different Frameworks

A common taxonomy for MWEs splits them first into lexicalized phrases and institutional phrases (Baldwin and Kim, 2010). Only the lexicalized phrases provide examples of syntactically deviant structures. They are in turn divided into fixed, semi-fixed, and syntactically flexible. This division can be seen as points on a scale from the most rigid to the fully compositional phrases (Masini, 2019). Here the focus will be on the fixed MWEs.

(Baldwin and Kim, 2010) defines fixed MWEs as expressions *'that undergo neither morphosyntactic variation nor internal modification, often due to fossilisation of what was once a compositional phrase.'* Expanding on this definition we have identified a number of ways in which a fixed MWE can vary, which is detailed in Section 3.

An interesting aspect of this definition is that it views fixed MWEs as isolated examples. Similarity of structure to other fixed MWEs seems to play little role. However, to determine whether an expression is fixed or flexible it is important to look for structural patterns that are common to sets of expressions, a key feature of Construction Grammar.

2.1. On Constructions

There are a number of variants of Construction Grammar but all of them use a notion of construction as a pairing of form and meaning. This applies to words and morphs as well as to phrases and clauses. The form level may include phonetic and/or orthographic information as well as morphological and syntactic information. Meaning may include semantic as well as pragmatic information (Hoffmann, 2022).

The morphosyntactic information is not restricted to parts-of-speech and morphological features. Depending on the scope of a construction the application of a category may be constrained in various ways, for instance to a subset of nouns or adjectives. Moreover, constructions are related to one another via inheritance links and horisontal links. In this way a phrase that seems deviant or special may be linked to a more regular pattern as a specialisation.

2.2. An Example

There is a set of Swedish time adverbials that are marked by the simultaneous occurrence of the preposition *i*, 'in' and a final suffix *-(a)s* on the following noun. The nouns are restricted to a finite number of words referring to week-days, seasons, or parts of the day. The sufix only occurs in this pattern. All expressions of the pattern are deictic and the meaning is, roughly, a reference to the most recent period of the kind signified by the noun:

i lördags	this past Saturday
i våras	this past spring
i julas	this past Christmas
i förmiddags	this past (late) morning

It is important to note that the nouns cannot be put in other nominal positions, not even as possessive modifiers. While *-s* is a genitive suffix in Swedish, the nouns in this group are seldom seen as possessive modifiers. For example, to say the equivalent of English 'the events of Saturday', in Swedish, we need to use a definite form, *lördagens händelser*, whereas an indefinite form such as **lördags händelser* on its own is out[1].

A construction in Usage Construction Grammar (Hoffmann, 2022) representing this set of time adverbials may be written as in Table 2.

Instances of this pattern that are found in Swedish UD treebanks are all annotated with the

[1] The label *kalenderplacering.genitiv*, 'calendar placement, genitive', which is found in the Swedish Constructicon (Borin et al., 2012; Lyngfelt et al., 2018) for these expressions is therefore unfortunate.

FORM:	[i NOUN$^1_{temp} + (a)s$]
MEANING:	this past $TIME^1$

Table 2: A construction in the style of a Usage CxG. The index links the noun in the FORM part to its corresponding predicate class in the MEANING part.

relation *fixed*. While there are only a finite number of them there is a clear pattern that capture their form as well as their meaning.

In a CxG patterns can be related to each other via inheritance, or as specifications of a common more general pattern. In the example we refer to more specific variables than ordinary parts-of-speech, such as week-days or seasons. This option is not available in UD, nor is UD concerned with meanings. However, a similar reasoning can be applied by relating the expression to a more general pattern captured by the part-of-speech variables ADP and NOUN. The normal relation assigned to an adposition in UD in front of a noun is *case* and the structure of the pattern can be captured as for other prepositional phrases as shown in Figure 1. Now we capture the syntactic structure of these expressions reasonably well. However, the information that we are dealing with a fixed expression has been lost. In the current UD framework we cannot say both at the same time. In the wording of (Gerdes and Kahane, 2016) the framework has created a catastrophe.

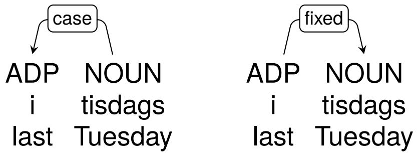


Figure 1: Two competing analyses of a fixed MWE, one as syntactically transparent and another as fixed.

Moreover, the pattern is similar to that of an adverbial expression consisting of a preposition and a non-inflected noun such as *på lördag* 'on Saturday', and *i morgon*, 'tomorrow'. Yet another similar structure employs rest morphemes such as *i går*, 'yesterday' and *i fjol*, 'last year'. Generalising further we can observe that other parts-of-speech such as adjectives can follow a preposition in expressions such as *inom kort*, 'shortly'. In UD we could view all of these as specializations of a common general pattern, ADP + ANY[2].

2.3. More on *fixed* in UD-treebanks

As stated in the introduction, *fixed* is only one of the three relations used for MWEs in UD. These relations have different properties, however. The *compound*-relation can go both to the left and the right and be embedded under a different *compound*-relation. This is not the case for *fixed* and *flat*; they are headless in principle but have the leftmost part as the head by default. Moreover, a dependent of *fixed* or *flat* can't have dependents of its own. Another UD relation with the same property is *goeswith*, which is primarily used for tokens that have been split accidentally. Structurally *fixed, flat* and *goeswith* can all be regarded as the same relation, just labelled differently for complementary information.

A special feature of *fixed*, according to its description on the UD web[3], is that it should be restricted to the most grammaticalized cases and be treated as a closed class. It is recommended that language-specific documentation is developed where the expressions for which *fixed* is applied are listed. The main reason for this is to enforce annotation consistency across treebanks in a way that can be validated automatically. This is definitely a worthy aim as the variation in its use is quite considerable. See Table 3 for figures on *fixed* in a sample of UD Treebanks, version 2.13. It can be noted that there are differences even for treebanks sharing the same language. In fact, some treebanks not shown in the table, like the Norwegian ones and UD_German-HDT do not use *fixed* at all. This shows that recommendations are motivated. It is likely that the differences are not due to language differences but to different annotation principles.

There are published lists only for a few languages, including English and Finnish. The English list has some 40 items, Finnish has around 90. The number of fixed expressions in the largest Finnish treebank is larger, however.

The idea to restrict fixed MWEs in UD to a smaller group raises the question how well it aligns with the notion of a fixed MWE as characterized in general works on the topic such as (Baldwin and Kim, 2010). Is it actually possible to find general criteria that could restrict the application of *fixed* in a principled way? This is investigated in Section 4.

3. Dataset and annotation

The main empirical data for the analysis are taken from the two Swedish UD treebanks UD_Swedish-Talbanken and UD_Swedish-Lines of version 2.13. In addition, we have looked at the list of proposed

[2] Instead of ANY we could specify a disjunction of UPOS categories.

[3] https://universaldependencies.org/u/dep/fixed.html

Treebank	Listed	In TB	%
UD_Dutch-Alpino	-	1161	2.75
UD_English-EWT	44	40	0.50
UD_English-GUM	44	44	0.64
UD_English-LinES	44	117	1.06
UD_Finnish-FTB	90	198	0.66
UD_Finnish-FTB	90	27	0.37
UD_French-Rhapsodie	-	70	2.62
UD_French-Sequoia	-	82	1.45
UD_Icelandic-IcePaHC	-	20	0.14
UD_Icelandic-Modern	-	2	0.05
UD_Italian-ISDT	-	79	0.66
UD_Italian-TWITTIRO	-	23	0.55
UD_Swedish-LinES	-	117	1.59
UD_Swedish-Talbanken	-	392	3.12

Table 3: Usage of *fixed* in a sample of UD treebanks. The column **In TB** shows the number of different types of MWE that are found in the treebank, while the column **%** shows the percentage of all tokens in the treebanks that carry *fixed* as their dependency.

English fixed expressions[4].

Together the two Swedish treebanks have 439 different MWEs annotated with *fixed*. Of these 71 are common to both treebanks, and 216 are hapaxes. For a few common MWEs, such as *som om*, 'as if', and *mer än*, 'more than' the two treebanks have made opposite decisions. Yet, the large majority satisfies the loose criterion of being multiword sequences that behave as function words, adverbs, or are special in some other way. As the treebanks are not very big we can safely assume that there are many more expressions that satisfy the same tolerant criteria as those in the treebanks. To compare, Wikipedia has 649 expressions listed under the label Swedish idioms and a recent dictionary of Swedish idioms (Luthman, 2020) contains 5000 items, although the majority of these are flexible.

Starting with the properties listed in the definition above (Baldwin and Kim, 2010) other properties were added as cases were found. Previous work on idioms in Swedish such as (Anward and Linell, 1976; Sköldberg, 2004) have largely focused on flexible idioms, but they define various criteria for recognizing MWEs including fixed expressions that we have considered. The expressions in the dataset have also been checked against larger Swedish corpora and concordances generated from the Korp interface[5] on news media. In the end we came up with 13 different properties as listed below. The first two relate to the expression's function and pattern, while the rest focus on some aspect of variation.

- **UPOS tag:** Part-of-speech if regarded as a single UD word, using the UPOS set of tags.
- **Syntactic pattern:** The syntactic pattern is expressed in terms of UPOS tags and regarded as the best generalisation of a more specific CxG pattern
- **Morpheme status:** Takes the values Roots, Inflected, Foreign, Abbr(eviation) and Special where Special includes rest morphemes and rare (obsolete) inflections.
- **Inflection variation:** Does any part of the expression allow inflectional variants? Yes or No.
- **Internal modification:** Does any part allow one or more modifiers? Yes or No.
- **Synonyms:** Is it possible to replace any part with synonyms? Yes or No.
- **Iterability:** Can a part be repeated? This is rare but occurs for several expressions that signify repeated events: *om och om (och om) igen*, 'again and again (and again)' Yes or No.
- **Order change:** Can the order among parts be different? Yes or No.
- **Optional part:** Is any part optional, or can an optional part be added? The answer is Yes or No and an example is *under det (att)*, 'while'.
- **Separability:** Can (or must) some part be separated from the rest by other material? Possible values are No, Obligatory, and Optional.
- **Idiom part:** Does the expression mainly occur as part of a longer idiom, in the treebank and generally? If so the value is Yes, otherwise No.
- **Abbreviation:** Does an abbreviated form exist? Yes or No.
- **Collapsibilty:** Does a single token equivalent exist? Often this is the result of omitting spaces as in *över allt : överallt*, 'everywhere'. Yes or No.

Every expression in the dataset has been described with these attributes. An illustration is given in Table 4 for the expression *i våras*[6]. Descriptions for the full dataset can be found in the supplementary material.

4. Types of fixed MWEs

Given the requirement that fixed expressions in UD should be a restricted closed class we want to

[4] https://universaldependencies.org/en/dep/fixed.html

[5] https://spraakbanken.gu.se/korp/

[6] In the expression *i fjol våras*, 'the spring of last year', we do not regard *fjol* as a modifier of *våras* but rather see it as a compound of two expressions *i fjol* and *(i) våras*

Attribute	Value	Comment
UPOS tag	ADV	
Pattern	ADP NOUN	
Morpheme status:	2:Special	*våras*
Inflection variation	No	
Modification	No	
Synonyms	No	
Iterability	No	
Order change	No	
Optional part	Yes	*i fjol våras*
Separability	No	
Idiom part	No	
Abbreviation	No	
Collapsible	No	

Table 4: Description of the Swedish expression *i våras*, 'this (past) spring' with respect to structure and variability.

reduce the number of expressions currently annotated with *fixed* in the Swedish treebanks. This entails two main things: identifying criteria that make *fixed* correspond well to a natural class of fixed expressions, and finding alternative dependency analyses for those expressions that are removed.

There are many different types of expressions in the dataset and the available space does not allow us to discuss all of them. We start with one type of variation that may be more common in a Swedish dataset than for other languages, the alternative renderings captured by the property of Collapsibility.

4.1. Collapsible MWEs

Swedish language planning authorities are generally quite tolerant towards variation in written Swedish. As a result many multiword expressions have alternative renderings as single tokens or, in case of three-part expressions, two tokens. As UD maintains that tokenisation should follow the orthographic rendering as far as possible, in particular that in-token spaces should be avoided, these expressions pose a special challenge.

In the dataset we find 75 collapsible MWES, which is about 17% of all. The large majority of them has an alternative rendering by omitting spaces. Examples are *till buds :: tillbuds*, 'at hand', *i dag :: idag*, 'today', *över huvud taget :: överhuvudtaget :: överhuvud taget*, 'actually'. The share of a certain rendering differs with individual expressions. We have investigated their distribution in two subsets of the Swedish Gigaword Corpus (Rødven Eide et al., 2016), news and fiction. The numbers support a division into three different groups, one where the the MWE rendering is much more common, one where the spaceless rendering is much more common, and one where the two renderings are about equally common. However, the relevance of this variation lies not so much in the exact proportions but that both renderings occur. A treebank should as far as possible assign the same analysis to both alternatives; they contain the same lexemes, but are just written differently. If spoken they would come out identical. Compare the two renderings below of the same sentence:

(3) *Hon kan när som helst komma i kapp*
(4) *Hon kan närsomhelst komma ikapp*
'She may catch up at any moment'

Given the aversion against token internal spaces in UD one option is to regard the multipart variants as basic and treat the shorter variants as multiword tokens. This solution aligns well with the long-term proposal for modelling synthetic compounds in UD put forward by (Savary et al., 2023b). A drawback is of course that this solution is sofar unseen in any Swedish treebank. Conversely, the existece of the single-token forms may be taken as an argument that they are perceived as single lexemes.

Using multiword tokens for the tokenisation of sentence (4) we would get the tokenisation in Table 5.

1	Hon	hon
2	kan	kunna
3-5	närsomhelst	_
3	när	när
4	som	som
5	helst	helst
6	komma	komma
7-8	ikapp	_
7	i	i
8	kapp	kapp

Table 5: Proposed tokenisation for single token alternatives to Swedish fixed MWEs.

4.2. Syntactic alternatives to *fixed*

For many of our expressions in the dataset we can find patterns that are shared with other expressions, as in Section 2.2. We may distinguish self-contained patterns from patterns with outward-looking parts. In the first type all included words except one have their head within the pattern. They are easy to provide a syntactic analysis for. With outward-looking parts two words have their heads outside of the pattern. Usually one of them is the last token which may be a preposition, subjunction or conjunction.

Self-contained expressions. The most common type of self-contained fixed expression in the dataset consists of a preposition followed by an uninflected noun. There are 66 such **prepositional**

phrases with examples such as *i dag*, 'today', *i allmänhet*, 'in general'. Other two-part expressions beginning with a preposition has a noun in definite form as head, *på vippen*, 'on the verge', an adjective, *på nytt*, 'anew', or a pronoun, *före detta*, 'ex-'. For some the UPOS is even hard to determine *på sistone*, 'lately', *på glänt*, 'slightly open', as the token is invariable and only occurs in this special expression. In addition there are three-part expressions with a nominal head of some sort. Taken together prepositional phrases account for almost 40% of the expressions in the dataset.

The syntactic structure of these prepositional phrases need not deviate from compositional phrases of the same patterns, see Figure 2. The fact that the correct UPOS tag for some words may be hard to determine does not prevent the assignment of an appropriate structure either. Moreover, the treatment of prepositions would actually be more consistent if they always are assigned the relation *case* when followed by a candidate head word.

We note that no more than four of the English MWEs in the list of English fixed MWEs are prepositional phrases, (*in order, of course, in case, at least*) and see this as support for treating prepositional phrases as non-fixed in the general case.

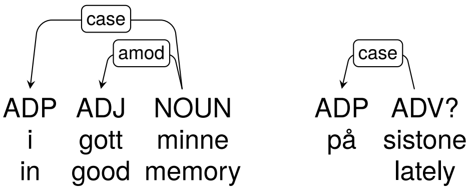


Figure 2: Syntactic dependency analysis for expressions beginning with a preposition.

Coordinations can be handled in the same way as prepositional phrases, since their syntactic structure is transparent when a coordinating conjunction is present. The most common type coordinates two adverbs but Swedish also shows instances of coordinated prepositions. Both structures can be viewed as specializations of a more general pattern for coordinations that need not require the two conjuncts to have the same part-of-speech. Thus, a fixed MWE as English *by and large* could be dealt with in the same way. The proposed structures are shown in Figure 3.

Another common type of pattern has an adverb or adjective as head modified by another adverb. Examples are *så pass (stor)*, 'that (big)' and *illa nog*, 'bad enough'. They also can be assigned the same structure as their compositional counterparts with the adverb serving as an *advmod*.

There are also expressions where an adverb seemingly modifies a preposition as in *in i*, 'into' or *fram till*, 'up to'. This is generally forbidden in the UD framework. To avoid annotating the adverb as a modifier we may regard the two parts as independently modifying the head.

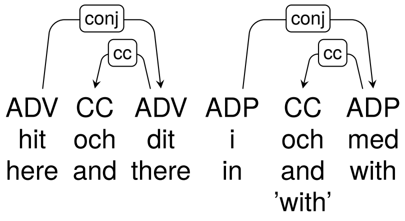


Figure 3: Syntactic dependency analysis for expressions employing coordinations.

Some of the expressions annotated with *fixed* end with a verb form of some sort most often a participle. Examples are *strängt taget*, 'actually', *allvarligt talat*, 'seriously speaking'. Regarded as verb phrases these expressions have obvious syntactic annotations: the participle is the head and the adverb an adverbial modifier. In relation to its context it may be annotated as an adverbial clause, *advcl*.

Outward-looking parts. A number of two- or three-word expressions have a last part that normally begins a phrase or clause of some sort. This applies to expressions ending in a preposition, a subjunction or one of the comparative conjunctions *än*, 'than' and *som*. 'as'.

The most common type of these are three-part sequences starting and ending with a preposition and a noun or nominal word in between. There are 48 expressions of this type in the dataset; examples are *på grund av*, 'because of', and *i samband med*, 'in connection with'.

Sometimes the final preposition introduces an optional phrase. An example is *med hjälp av*, 'with the aid of', where *med hjälp* can act as an adverbial phrase on its own. In those cases it is perfectly reasonable to view the noun in the middle as the head. See Figure 4. If the preposition is required, however, as in *på grund av*, 'because of', this solution can be questioned. We note though that in the English list of *fixed* expressions, this type of three-part expression is rare. For example, *in spite of* is not included so that *spite* comes out as the head of a noun phrase such as *in spite of the problems* giving the same structure as in Figure 4[7]

Expressions ending with a subjunction are also quite common; in the data set we find 9 ending in *att*, 'that', 2 ending in *om*, 'if', and 10 ending in *som*, 'as'. Here a different analysis may be advocated: assigning the different parts separate functions as

[7] For example, the tree with sent-id 'weblog-blogspot.com_alaindewitt_20060924104100_ENG_20060924_104100-0031 in en_ewt-ud-train.conllu

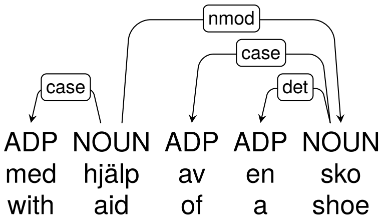


Figure 4: Syntactic relations for the three-part expression *med hjälp av*, 'with the aid of'.

mark or *case* depending on the part-of-speech. For example, in the case of *som om* and, similarly, *as if*, one may argue that each of the two parts has a function of its own. The first, *som/as* indicates that we are dealing with a comparison, the second, *om/if* that we are dealing with something unreal or assumed. In Swedish, such an analysis gains some support from the fact that the *if*-clause in certain circumstances can be replaced by a clause without the subjunction:

(5) *Han beter sig som vore han ...*
'He behaves as were he ...

(6) *Han uppför sig som om han var ...*
'He behaves as if he were ...

There are eight expressions ending with the comparative conjunction *än*, 'than'. The majority are introduced by an adjective or adverb in comparative form, such as *mer än*, 'more than', *lägre än*, 'lower than' or 'less than'. The comparatives actually all accept modifiers such as *mycket*, 'much', or *lite*, 'a little', and for this reason they may not qualify as fixed expressions. Syntactically they can be treated as other expressions with outward-looking parts, letting the conjunction find its head to the right and the whole of that complex be a dependent to the word in the comparative.

In the English treebanks the expressions *more than* and *less than* are regarded as fixed when they modify a quantity as in *more than 90 percent* but not in other contexts. This is a bit awkward as there is no difference in the possibility of adding the modifier *much*: *much more than I have* and *much more than 90 percent* sound equally well-formed.

Similar arguments apply to comparison using the conjunction *som*, 'as'. They are common both in our dataset and in the English list. But they often share a pattern as the English *as many/much/few/little as* where virtually any adjective and a number of adverbs may occur in the middle. This indicates that we are dealing with a construction that can be annotated as such with the adjective/adverb as the head.

4.3. Types based on variation

Another basis for grouping expressions is the amount of variation that they admit. For our dataset we may distinguish three groups. At one end there are expressions with no or almost no variation based on the variational attributes that may be called **rigid**. At the other end we find several expressions that allow inflectional variation, replacement with synonyms and/or internal modification. Those will be called **semi-flexible**.

Semi-flexible expressions. 57 of the expressions that are currently annotated with the relation *fixed* can actually be varied enough to be called semi-flexible. This applies to expressions with parts that can be inflected in accordance with their part of speech, be replaced by synonyms, and/or take modifiers. Expressions of this type are

- *när det gäller*, 'concerning', (inflectional alternatives *gällt, gällde*, other alternative *vad det gäller*.
- *vem som helst*, 'whoever', (modifiers *fan*, 'the devil', *av dem*, 'of them', and similarly for other expressions of the same pattern: *när som helst, var som helst*, 'whenever', 'wherever'.
- *den här*, 'this', *den där*, 'that'. with variants *de, den, det, dom* for the first part, and *här, där* for the second part. The second parts are also found after *så, sådan, sådant, sådana* giving expressions meaning 'like this' or 'like that'.

For these types we argue that they shouldn't be regarded as fixed MWEs at all because of the amount of variation they accept. Instead syntactic analyses need to be found.

Rigid expressions There are 96 expressions in the dataset that show no variation at all. By including those that are collapsible and/or have an abbreviated form we reach 146 expressions. The most common are *som om*, 'as if', *så att*, 'so that', *i dag*, 'today', *därför att*, 'because', *på grund av*, 'because of', *för att*', '(in order) to', *i stället*', 'instead', *till exempel*', 'for example', all of which occur more than 30 times in the treebanks. We note that in case the English counterparts are MWEs they are listed as *fixed* for English[8]. Rigidity may thus be regarded as a characteristic property of expressions to be annotated as *fixed*.

Also included in this group are expressions from other languages and abbreviations. They are not so numerous but illustrate general types of interest.

[8] In the case of *in order to*, however, only *order* is taken as a dependent of *in*, while *to* finds its head in a verb to the right

There are expressions of Latin origin such as *a priori* and *vice versa* and one of English origin, *to date*. Abbreviations include short forms of academic degrees such as *med lic*, 'licentiate in medicin' and common phenomena in academic prose, such as *a. a.*, short for 'anfört arbete', and a counterpart to the Latin 'op. cit.'.

In UD foreign material may be annotated in different ways. If regarded as a borrowing it should be given a suitable UPOS tag and different parts be connected via the relation *flat* (sic!). If regarded as truly foreign each part should have the UPOS X, and, in addition, carry the feature information FOREIGN=Yes. The parts should again be connected via *flat*. With one exception, the expression *ad calendas graecas*, the examples in the dataset are sufficiently common in Swedish to be regarded as borrowings. Depending on their status as functional (*vice versa*) or not (*ad hoc* they could fit either *fixed* or *flat*.

Abbreviations should be marked by the feature Abbr=Yes. The UPOS tag should reflect the part-of-speech of the abbreviated word. The expanded versions of our two examples both consist of an adjective and a noun so the dependency analysis could use the *amod*-relation rather than *fixed*. See Figure 5.

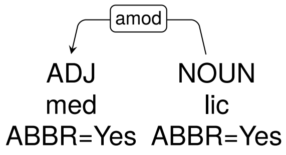


Figure 5: Dependency analysis of the abbreviated title *med lic*, 'licentiate in medicine'.

4.4. Candidates for the *fixed* list

. A large number of MWEs currently marked as *fixed*can be excluded as candidates for the list of *fixed* expressions on the basis of their morphosyntactic variation. With a fairly strict criterion on rigidity, not excluding MWEs that are collapsible or can be abbreviated, there are 146 items left. By considering that *fixed* should be restricted to items with function word distribution another seven can also be removed, leaving 139. This is still a large number, however, especially considering that the treebanks only cover a subset of the Swedish MWEs. On the other hand, many of them have a transparent syntactic structure; being self-contained expressions of the kinds described in Section 4.2. By consistently preferring a headed structure when the MWE satisfies such a pattern the numbers can be reduced further. Other types that may be excluded are those where different parts of the MWE can be separately annotated with a dependency to an outside head as was argued in the case of *som om*, 'as if' and as is done in English treebanks with many MWEs of the form 'ADP NOUN ADP'.

As UD is reluctant to see function words as heads the most likely MWEs to put on the list of items annotated with *fixed* are two-word MWEs ending in a preposition or a subjunction. Examples of the first kind are such *in i*, 'into' and *rent av*, 'actually' and of the second *så att*, 'so that', *för att*, '(in order) to', and *ifråga om*, 'as regards'. Another set of likely candidates come from adverbial and prepositional MWEs where the head word is not an adverb or a preposition as for *tack vare*, 'because of', *till synes*, 'seemingly'.

5. Alternative annotations of fixed expressions in UD

The current UD guidelines on fixed expressions hide their, in many cases, apparent syntactic structure. (Gerdes and Kahane, 2016) have pointed out this as a 'catastrophe' problem and makes a proposal to subcategorize syntactic dependencies with a special identifier such as *mwe*. A disadvantage of this solution is that it will profilerate the *mwe* subcategory in the trees. Moreover it annotates the property of being a multiword expression at a single level to the exclusion of other properties that an MWE may have. The proposal in (Kahane et al., 2017) to insert extra lines for fixed expressions such as *top of the range*, which may carry a dependency relation of its own seems more accurate for capturing the lexical character of fixed expressions.

An alternative is to unify the shallow headless relations to one, say *flat*[9], and treat a property such as fixedness with a feature in the same way as is done with foreignness. This would make the annotation similar to that for split words, where the relation *goeswith* is used in tandem with the feature Typo=Yes[10]. The features for a fixed MWE could then be applied to its head and be interpreted as including the dependents by default.

This solution would also solve the problem of choosing between *fixed* and *flat*. As shown above the properties of phrases as being fixed, abbreviated, or from a different language sometimes converge. An expression such as *vice versa* could actually be annotated as foreign and fixed at the same time. Then the *fixed* is in conflict with *flat* which is recommended for foreign material. Annotating these properties at the level of features allows them to be combined.

[9] A similar proposal is made in (Savary et al., 2023b) using the label *headless*.

[10] `https://universaldependencies.org/u/dep/goeswith.html`

A third more radical alternative is not to deal with fixed expressions at all in the current UD format. While there is a need to mark headlessness in the syntactic trees, it is evident that not all kinds of MWEs can be handled as part of UD dependency trees. It is also evident that the current feature annotation is insufficient. It is restricted to words and thus cannot cover subtrees with one feature. The CUPT format (CoNLL-U Plus Format) as used by the PARSEME:MWE framework for annotating verbal MWEs allows more complex feature annotation and may be used for many types of MWEs including fixed expressions. This seems to be the future that is also envisioned by (Savary et al., 2023b).

With this alternative appropriate syntactic dependencies need to be found. We have suggested that a Construction Grammar perspective on fixed MWEs is helpful for this purpose. UD has a general principle of a tight relation between UPOS categories and dependency relations. This principle could be extended to UPOS sequences that share enough common features to be related hierarchically to a dependency template as suggested in Section 2.2.

6. Conclusions

We have analysed 439 expressions currently annotated as fixed expressions in Swedish UD treebanks with the aim of producing a well-defined subset that meets UD requirements on the use of the relation *fixed*. We have found a way to reduce this set by closely studying their variational properties and the structural patterns that they share. Although we find a number of rigid MWEs, i.e., expressions admitting no or almost no variation at all, they often have a transparent syntactic structure which is not accounted for when *fixed* is used. And many of them share structure with other MWEs. These structures can be represented in more detail in Construction Grammar frameworks, as we have shown with examples. Although UD does not allow such detail we can nevertheless often generalise the structure to something that can be expressed in UD-terms. Moreover, to capture all kinds of MWEs, whether fixed or flexible, requires a more versatile format than CoNLL-U such as the CUSP-format used for annotating verbal MWEs in the PARSEME:MWE project.

Annotating fixed expressions with a specific relation as part of the dependency structure, as is currently done in UD, prevents the annotation of its syntactic structure. A better solution would be to isolate the structural properties of *fixed*, which it shares with other UD relations such as *flat* and *goeswith*, in a single relation and use features to indicate the character of the expression, something which now is done only for typos.

Another problem we discovered, which may be specific to Swedish, is the large numbers of collapsible MWEs. The best solution we could propose for these, in order to ensure that the dependency analysis would come out the same whether the MWE is collapsed or not is to make use of UD's provision of multiword tokens.

7. Acknowledgements

I am indepted to Joakim Nivre for valuable comments on an earlier version of the paper and to an anonymous reviewer for detailed critique and pointing me to the article (Savary et al., 2023b). The work has been partially funded by the Swedish Research Council through the Swe-Clarin part of the Swedish National Language Bank.

8. Optional Supplementary Materials

A spreadsheet with our analysis of 439 MWEs currently analysed as fixed in Swedish treebanks is provided as supplementary material.

8.1. Extra space for ethical considerations and limitations

This work is based on open resources and, as far as we can see, pose no ethical problems. A limitation is that it is based on treebank data from one language only and some comparisons with English data. We are certain, though, that the types of problematic multiword expressions discussed here can be found also in other UD treebanks. However, the restriction to one language means that the list of types is likely to be incomplete.

9. Bibliographical References

Jan Anward and Per Linell. 1976. Om lexikaliserade fraser i svenskan (lexicalized phrases in Swedish). *Nysvenska studier. (Studies in Modern Swedish)*, 55/56:77–119.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Boca Raton, USA,.

Geert Booij. 2017. Construction morphology. In Mark Aronoff, editor, *Oxford research encyclopedia of linguistics.* Oxford University Press, New York.

Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Julia Prentice, Rudolf Rydstedt, Emma Sköldberg, and Sofia Tingsell. 2012. Growing a Swedish constructicon in lexical soil. In *Proceedings of SLTC 2012. (The Fourth Swedish Language Technology Conference)*, pages 10–11.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

Kim Gerdes and Sylvain Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of Universal Dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 131–140, Berlin, Germany. Association for Computational Linguistics.

Thomas Hoffmann. 2022. *Construction Grammar: The Structure of English*. Cambridge University Press.

Sylvain Kahane, Marine Courtin, and Kim Gerdes. 2017. Multi-word annotation in syntactic treebanks: Propositions for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 181–189, Prague, Czech Republic.

Hans Luthman. 2020. *Svenska idiom: 5000 vardagsuttryck (Swedish idioms: 5000 everyday expressions)*. Folkuniversitetets förlag.

Benjamin Lyngfelt. 2021. Valens och konstruktioner - om samspelet mellan lexikon och grammatik (valency and constructions - on the interplay of lexicon and grammar. In Johan Brandtler and Mikael Kalm, editors, *Nyanser av grammatik. Gränser, mångfald, fördjupning*. Studentlitteratur.

Benjamin Lyngfelt, Linnéa Bäckström, Lars Borin, Anna Ehrlemark, and Rudolf Rydstedt. 2018. Constructicography at work: Theory meets practice in the Swedish constructicon. In B. Lyngfelt, L. Borin, K. Ohara, and T. T. Torrent, editors, *Constructicography: Constructicon development across languages*, pages 41–106. John Benjamins, Amsterdam.

Francesca Masini. 2019. Multi-word expressions and morphology. In Mark Aronoff, editor, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press,.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish Culturomics Gigaword Corpus: A one billion word Swedish reference dataset for nlp. In *Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts*, pages 3–36, Krakow, Poland. John Benjamins Publishing Company.

Ivan A. Sag, Timothy Baldwin, Frances Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, and et al. 2023a. Parseme corpus release 1.3. In *Proceedings of The 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023b. Parseme meets universal dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9(1).

Emma Sköldberg. 2004. *Korten på bordet: Innehålls- och uttrycksmässig variation hos svenska idiom (Cards on the table. Variations in content and expression in Swedish idioms)*. Ph.D. thesis, University of Gothenburg.

Synthetic-Error Augmented Parsing of Swedish as a Second Language: Experiments with Word Order

Arianna Masciolini, Emilie Marie Carreau Francis, Maria Irena Szawerna
Språkbanken Text
Department of Swedish, Multilingualism, Language Technology
University of Gothenburg
{arianna.masciolini, emilie.francis, maria.szawerna}@gu.se

Abstract

Ungrammatical text poses significant challenges for off-the-shelf dependency parsers. In this paper, we explore the effectiveness of using synthetic data to improve performance on essays written by learners of Swedish as a second language. Due to their relevance and ease of annotation, we restrict our initial experiments to word order errors. To do that, we build a corrupted version of the standard Swedish Universal Dependencies (UD) treebank Talbanken, mimicking the error patterns and frequency distributions observed in the Swedish Learner Language (SweLL) corpus. We then use the MaChAmp (Massive Choice, Ample tasks) toolkit to train an array of BERT-based dependency parsers, fine-tuning on different combinations of original and corrupted data. We evaluate the resulting models not only on their respective test sets but also, most importantly, on a smaller collection of sentence-correction pairs derived from SweLL. Results show small but significant performance improvements on the target domain, with minimal decline on normative data.

Keywords: Dependency Parsing, Data Augmentation, Second Language Acquisition, L2 Swedish

1. Introduction and Background

In recent years, off-the-shelf dependency parsers have reached remarkably high performance on standard evaluation sets. This applies to many high and medium-resourced languages, including Swedish. Nonstandard language, however, still poses significant challenges. In a study on dependency parsing of learner English, Huang et al. (2018) showed that the tools available at the time were not robust to grammatical errors, despite misleadingly high overall accuracy scores. In a more recent study on L2 Swedish (Swedish as a second language), Volodina et al. (2022) note that, dependency parsing is especially problematic for standard tools, even when they perform reasonably well on other linguistic annotation tasks such as part-of-speech tagging.

A notable attempt to address this issue is the error-repairing parser introduced by Sakaguchi et al. (2017), specifically meant for ungrammatical texts. This approach combines parsing with Grammatical Error Correction (GEC). In many contexts, such as Second Language Acquisition (SLA) research, it can however be preferable to analyze learner texts as they are and, in some cases, to compare originals with their normalized versions. We therefore test the more straightforward approach of fine-tuning a Bidirectional Encoder Representations from Transformers (BERT, Devlin et al. 2018) model for dependency parsing on data that resembles our target domain, L2 Swedish.

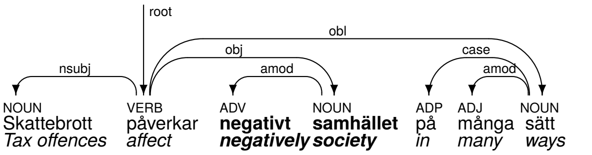


Figure 1: A sentence with incorrect word order parsed with UDPipe 2. Note how the adverb *negativt* is both attached to the wrong token (the noun *samhället*, rather than the main verb *påverkar*) and incorrectly labelled as an adjectival modifier (`amod`) instead of as an adverbial one (`advmod`).

With an approach loosely inspired by Stymne et al. (2023), we use the MaChAmp (Massive Choice, Ample tasks) toolkit (van der Goot et al., 2021) to fine-tune an array of models on different combinations of a treebank of standard Swedish and an artificially corrupted version of the same dataset. Crucially, the evaluation step involves not only normative data and artificial errors, but also authentic L2 Swedish sentences.

For this first experiment, we restrict ourselves to word order errors. This is out of both principled and practical reasons. On the one hand, as illustrated by the example in Figure 1, it seems reasonable to assume syntax errors to be challenging for a tool that performs syntactic analysis. When it comes to word order errors specifically, this should be especially true for a language with relatively strict word order such as Swedish. At the same time, word order errors appear to be easier to generate and automatically annotate than most other error types: as tokens are swapped without being altered, token-level linguistic annotation can be easily transferred from a sentence in standard language to its corresponding corrupted version.

2. Data

We utilize three datasets: an L2 Swedish test set, described in Section 2.1, a standard Swedish treebank and an artificially corrupted version of the latter (cf. Section 2.2). Train-dev-test split sizes are outlined in Table 1.

2.1. SweLL

Our target domain data comes from the SweLL Swedish Learner Language corpus (Volodina et al., 2019), a collection of over 500 essays written by learners of L2 Swedish. More specifically, we use SweLL-gold, the manually pseudonymized version of the corpus (Volodina et al., 2022).[1] L1 backgrounds vary, as well proficiency levels, which range from beginner to advanced. Learner texts are paired with correction hypotheses[2] and each error is classified according to the taxonomy discussed in Rudebeck and Sundberg (2021).

For our purposes, the relevant categories are, in decreasing order of frequency, S-Adv (misplaced adverbial), S-FinV (misplaced finite verb), and S-WO, which encompasses all other word order errors. About 15% of SweLL sentences are marked with one of these labels. In the vast majority of the cases, however, word order errors co-occur with other issues, often overlapping in ways that make the former hard to isolate. After filtering out these cases, we were left with a 69-sentence evaluation set. Regrettably, the resulting sentences tend to be shorter than the corpus-wide average.

2.1.1. Linguistic Annotation

While a linguistically annotated version of SweLL is available, it is not manually validated nor does it follow the UD standard. We therefore opted for completely re-annotating our test set. We started by parsing the correction hypotheses with the UDPipe 2 parser (Straka, 2018) using the UD 2.12 model (Straka, 2023) trained on Talbanken (cf. Section 2.2). The first and third authors, both graduate students in Computational Linguistics, manually validated the resulting parse trees with particular attention to the segments that diverged from the corresponding original learner sentences. This manual annotation step only concerned the DEPREL and HEAD columns of the fully-annotated CoNLL-U files obtained from UDPipe 2, as our models are only trained for UD parsing in its strictest sense.

To annotate L2 originals, we used an *ad-hoc* script which transfers token-level annotations from gold-annotated corrections to L2 originals. Each sentence is first rewritten in the vertical format customary for CoNNL-U files. Then, each token is annotated as follows:

- a token ID is assigned sequentially;
- all other fields excepts HEAD (syntactic head) are copied from the first unused token of the sentence's correction hypothesis presenting the same word FORM. Such token is then immediately marked as used, to deal with cases where the same word occurs multiple times in the same sentence;
- the HEAD field is assigned the ID of the nearest token in the learner sentence whose FORM matches that of the syntactic head of the corresponding corrected token.

Choosing syntactic heads based on the closest homograph is a heuristic that occasionally produces ill-formed trees. For this reason, we also inspected the results of this processing step and made the necessary manual edits.

2.2. Talbanken

For training, we used the UD 2.12 version of Talbanken, a widely used treebank of written and spoken modern Swedish (Einarsson 1976, Nivre and Smith 2023). Due to MaChAmp not supporting the enhanced UD format, the treebank was preprocessed with the cleanup script provided as part of the toolkit itself. Its training portion was then used to fit our baseline model with no further changes. Mimicking the error patterns observed in SweLL, we also built a corrupted version of such a treebank, which we used in conjunction with the original upon training our specialized models (cf. Section 3).

2.2.1. Corruption Process

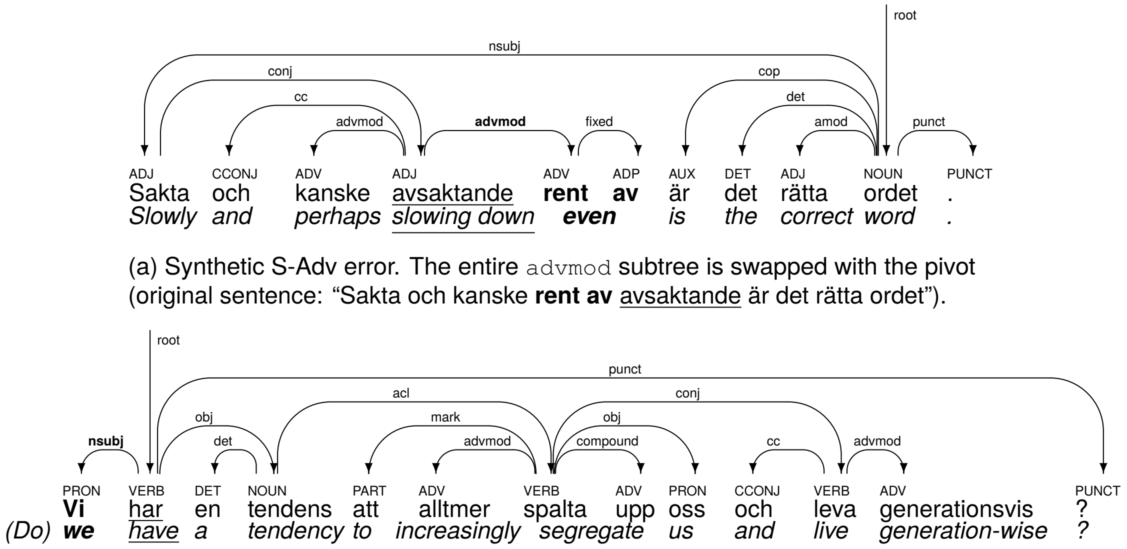
Synthetic error generation is a common task in the field of GEC. Closest to this work is the text corruption method described in Casademont Moner and Volodina (2022), which has been used to build a corpus of Swedish sentences presenting verb

	Train	Dev	Test
SweLL	-	-	69
Talbanken	4303	504	1219
Corrupted	4303	504	1219

Table 1: Sizes of the training, development and test splits of our datasets in number of sentences.

[1] For conciseness, we refer to SweLL-gold as SweLL.

[2] Annotators often need to guess the learner's communicative intent. For this reason, we refer to normalized sentences as correction hypotheses.


root
nsubj
conj
cop
cc
det
advmod
advmod
fixed
amod
punct
ADJ CCONJ ADV ADJ ADV ADP AUX DET ADJ NOUN PUNCT
Sakta och kanske avsaktande rent av är det rätta ordet .
Slowly and perhaps slowing down even is the correct word .
root
punct
acl
conj
obj
mark
obj
nsubj
det
advmod
compound
cc
advmod
PRON VERB DET NOUN PART ADV VERB ADV PRON CCONJ VERB ADV PUNCT
Vi har en tendens att alltmer spalta upp oss och leva generationsvis ?
(Do) we have a tendency to increasingly segregate us and live generation-wise ?

(a) Synthetic S-Adv error. The entire `advmod` subtree is swapped with the pivot (original sentence: "Sakta och kanske **rent av** avsaktande är det rätta ordet").

(b) Synthetic S-FinV error. Note that Swedish features reversed word order in questions (original sentence: "Har **vi** en tendens att alltmer spalta upp oss och leva generationsvis?").

Figure 2: Two corrupted sentences obtained via subtree swapping. The rearranged segments are higlighted in bold; their syntactic heads, acting as pivot elements, are underlined.

order errors using L2 Swedish textbooks as a starting point. We propose a simpler but more general method that covers all three classes of word order errors mentioned in Section 2.1 while preserving UD annotation.

From an operational point of view, such an approach resembles that of Şahin and Steedman (2018), who rely on dependency annotation to "rotate" sentences by swapping subtrees around roots. When it comes to misplaced adverbials (S-Adv), subtrees labelled as adverbial modifiers (`advmod`) or clauses (`advcl`) are swapped with their syntactic heads (see Figure 2a for an example). S-FinV errors are generated by swapping finite verbs with their subjects (a `nsubj`- or `csubj`-labelled subtree[3], cf. Figure 2b). As for S-WO, with a drastic simplification, we always swap two randomly selected adjacent tokens. After each rotation, the `ID`s of the corrupted sentence are reassigned sequentially and dependency `HEAD`s adjusted accordingly, thus ensuring the correctness of the annotation for the resulting corrupted tree.

We tried as much as possible to replicate the error distribution observed in SweLL. For each Talbanken sentence, our corruption script tries to generate three different scrambled sentences (one per error category) and chooses one based on its label's relative frequency in the corpus. Obviously, however, the S-Adv corruption rule cannot be applied to sentences with no adverbials. There are also instances where finite verbs (typically imperatives) lack an explicit subject or, more rarely, where sentences contain no finite verbs at all. In both cases, we revert to one of the other two categories.

3. Models

Name		% Normative	% Errors
BASELINE		100	0
MIX15		85	15
MIX50		50	50
SEQ10	(step 1)	100	0
	(step 2)	0	100
SEQ20	(step 1)	100	0
	(step 2)	0	100

Table 2: Our models and the data configurations they were trained on.

We used the MaChAmp toolkit to fine-tune a BERT model for dependency parsing using the original and corrupted Talbanken datasets in different configurations, summarized in Table 2. MaChAmp simplifies the fine-tuning of language models for a variety of NLP tasks including dependency parsing (van der Goot et al., 2021). It is relatively simple to set up with the desired hyperparameters and allows for the fine-tuning of various contextualized word embeddings. While we do not leverage the toolkit's multi-task learning functionalities, we have selected it for its ease of use and sequential fine-tuning. We ran the toolkit with the default hyperparameters, with the exception of changing the default model to the monolingual Swedish BERT (Malmsten et al., 2020) and altering the number of epochs in one

[3] If the finite verb in question is an auxiliary, we look for the subject of the head lexical verb.

of our sequential models (seq10 was only further fine-tuned on 10 epochs of corrupted data, not the default 20).

All in all, we have fine-tuned BERT for Swedish five times, resulting in five final models. The first model we fine-tuned purely on Talbanken, as a baseline (baseline), in order to know what results fine-tuning only on normative data yields. Our first specialized model, mix15, utilized a combination of normative data and synthetic errors that was meant to mimic the relative frequency of this kind of errors in the learner data. In order to see whether increasing that relative frequency would have a detrimental effect on a model, we fine-tuned one with equal parts of normative and corrupted data, mix50. We also experimented with sequential training to further fine-tune the baseline model with 10 and 20 epochs of only corrupted data (seq10 and seq20, respectively), to investigate whether the performance of an existing dependency parser could be improved by retraining it on non-normative language.

4. Evaluation

Model accuracy was evaluated in terms of Labelled and Unlabelled Attachment Scores, LAS and UAS. To check for statistical significance, these were calculated for each parse tree and compared against a baseline trained on standard Talbanken data to determine if the difference in model performance was significant. A paired t-test with a 95% confidence interval and $\alpha = 0.05$ was used with the Bonferroni correction to compensate for multiple tests against the baseline. Both the UAS score and LAS score were tested against the baseline, so it is possible for only one of the scores to be statistically significant. For nearly all cases, with the exception of Seq20 SweLL (Table 4), either both scores or neither were found to be significant.

Performance on target domain data was assessed on the SweLL-derived test set described in Section 2.1. The models were also evaluated on the original Talbanken test set and its corrupted version (cf. Section 2.2). Talbanken was included to assess whether the addition of ungrammatical examples resulted in a performance decline on normative data, while SweLL allowed for comparison of results on actual learner errors. The expectation was to see a substantial performance increase on corrupted Talbanken instances and a smaller improvement on authentic examples. When it comes to normative data, the ideal outcome would be for the fine-tuning on artificial errors to not have any negative repercussions.

Targeted Evaluation To further analyse how this method affects word order errors, a more targeted evaluation was performed using a modified version of the SweLL test set. Following Berzak et al. (2016), we assumed tokens belonging to erroneous segments to be more likely to be incorrectly parsed, even though annotation errors might cascade to other parts of the sentences. Errors were isolated from learner sentence-correction pairs by removing tokens preceding and following the diverging segment. Attachment scores were then recomputed on the resulting sentence fragments.[4]

4.1. Results and Discussion

Overall average scores are summarized in Table 3. Performance results suggest that exposure to synthetic word order errors in training has a positive effect on the models' ability to handle the (in-domain) corruputed sentences, matching our expectations. Simultaneously, performance decline on normative data is contained. Addressing the central question of whether improvement on synthetic data transfers to actual learner sentences, a slight positive effect on similar errors in out-of-domain texts can be observed. Smaller performance gains on out-of-domain texts may be attributed to synthetic errors not being sufficiently similar to authentic examples, to differences between training and test domains beyond mere grammaticality, or a combination of the two. It must also be taken into account that the margin of improvement on learner sentences is smaller than on artificial errors. On artificially corrupted sentences, the baseline's performance

	Talbanken		Corrupted		SweLL	
	LAS	UAS	LAS	UAS	LAS	UAS
baseline	92.42	94.30	80.20	83.29	88.28	91.16
mix15	92.23	94.05	87.96	90.50	87.63	90.60
mix50	91.54	93.58	89.59	92.00	89.86	92.93
seq10	92.20	94.06	90.47	92.75	90.05	92.84
seq20	92.53	94.32	90.95	93.08	89.02	92.00

Table 3: Overall attachment scores sets for all fine-tuned models. Cells with a grey background indicate that the difference between the scores for the baseline and fine-tuned models is statistically significant.

[4] Postprocessing often result in ill-formed trees, but this does not affect either performance metric.

drops by about 10% for both metrics, while scores stay reasonably high on SweLL. Notably, on the other hand, specialized models perform very similarly on both non-normative datasets. The SEQ10 model performed best across all test sets except Talbanken.

4.1.1. Talbanken

The Talbanken set showed the highest performance overall, with the baseline achieving a LAS of 92.42% and an UAS of 94.3%. This observation is expected, as the models were for the most part trained on the same domain (Talbanken data). Performance with the fine-tuned models generally decreased, but only MIX50 and SEQ10 showed a result that was significantly different compared to the baseline. It appears that exposing the model to atypical word order has little impact on performance for the Talbanken domain.

4.1.2. Corrupted Talbanken

Results for the corrupted Talbanken set showed the largest increase in performance compared to the baseline, about an 8 to 10% increase, and the differences were statistically significant.[5] The SEQ10 and SEQ20 models showed the biggest improvement, with a 10% increase over the baseline. This confirms the viability of the fine-tuning approach for specialized UD parsers, at least when target domain data is available.

4.1.3. SweLL

Most specialized models exhibited small performance improvements against the baseline. However, just the SEQ10 model's improvement was significant. Interestingly, the only model that declined in performance, MIX15, was the one exposed to a percentage of errors corresponding to the one observed in SweLL-gold, which appears not to be enough to produce a positive effect.

A further encouraging signal comes from the targeted evaluation. When we focus on ungrammatical fragments, we see that the performance gap between the baseline and all the specialized models widens (cf. Table 4). Not only does this confirm the baseline's vulnerability to grammatical errors, but it also suggests that the models are learning something about non-normative word order, rather than just exhibiting a general improvement due to exposure to additional training data.

5. Conclusions and Future Work

We generated synthetic word order errors and used them to fine-tune a number of dependency parsers. We evaluated them on (1) normative data, (2) synthetic error data, and (3) authentic L2 sentences containing errors of the same kind. The improvement on the latter was small, but significant. No substantial decrease in performance on normative data was observed, which suggests this is a promising method to increase parser robustness.

	LAS	UAS
BASELINE	82.80	86.02
MIX15	84.41	89.25
MIX50	87.10	90.32
SEQ10	87.10	89.78
SEQ20	86.02	89.78

Table 4: Attachment scores for the targeted evaluation on the SweLL-based test set. Cells with a grey background indicate that the difference between the scores for the baseline and fine-tuned models is statistically significant.

Future work aimed at achieving a more significant performance increase on target domain data should revolve around improving the corruption pipeline, especially when it comes to S-WO errors. The choice of material to corrupt is also important. In fact, we believe that applying our method to sentences from a domain closer to learner essays could result in better performance. It would also be beneficial to either have a larger test set or compare models in terms of multi-run averages in the future in order to more confidently assert that the differences between fine-tuning methods are not accidental. Other interesting possibilities are trying to run a hyperparameter search for at least some of the models and seeing how a multilingual model compares to the monolingual one we employed.

To ensure that our method is actually applicable to learner data in a more general sense, a possibility is to add one more test set where word order errors co-occur with other issues. Finally, a central question is to what extent our approach can be generalized to handle other kinds of errors (such as missing or redundant tokens, lack of agreement, etc.), and, most importantly, whether it can be adapted to handle sentences with multiple errors of various kinds.

6. Data and Code

The SweLL-derived test set and code are available at github.com/spraakbanken/seapass.

7. Ethical Concerns

While linguistic data can contain personal information, raising privacy concerns, neither of the datasets used in this experiment is likely to leak sen-

[5] p=0.0000000000000022, per paired t-test.

sitive information. Aside from its age, Talbanken consists of texts from genres like textbooks and articles, which are unlikely to contain information that should not be shared. As for SweLL-gold, a corpus that is both more recent and more likely to contain sensitive information due to its domain (L2 learner essays), all of the elements considered to be sensitive have been replaced with pseudonyms during corpus creation, and appropriate written consent had been obtained during the data collection step. Therefore, we consider the privacy risks of using these two datasets to be minimal.

8. Acknowledgments

SweLL-gold is part of Språkbanken, the Swedish national research infrastructure. Furthermore, the experiments presented in this paper are preliminary to the release of a UD version of such corpus, which will in turn enrich the existing infrastructure. On this basis, this research is supported by Nationella Språkbanken, funded jointly by the Swedish Research Council (2018–2024, contact 2017-00626) and the ten participating partner institutions.

9. Bibliographical References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

Judit Casademont Moner and Elena Volodina. 2022. Generation of synthetic error data of verb order errors for Swedish. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 33–38, Seattle, Washington. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jan Einarsson. 1976. *Talbankens skriftspråkskonkordans*. Institutionen för nordiska språk, Lunds universitet.

Yan Huang, Akira Murakami, Theodora Alexopoulou, and Anna Korhonen. 2018. Dependency parsing of learner English. *International Journal of Corpus Linguistics*, 23(1):28–54.

Lisa Rudebeck and Gunlög Sundberg. 2021. SweLL correction annotation guidelines. In *The SweLL guideline series nr 4*, Gothenburg, Sweden. Institutionen för svenska, Göteborgs Universitet.

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009, Brussels, Belgium. Association for Computational Linguistics.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2017. Error-repair dependency parsing for ungrammatical texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–195, Vancouver, Canada. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Sara Stymne, Carin Östman, and David Håkansson. 2023. Parser evaluation for analyzing Swedish 19th-20th century literature. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 335–346, Tórshavn, Faroe Islands. University of Tartu Library.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Elena Volodina, David Alfter, Therese Lindström Tiedemann, Maisa Susanna Lauriala, and Daniela Helena Piipponen. 2022. Reliability of automatic linguistic annotation: native vs nonnative texts. In *Selected papers from the CLARIN Annual Conference 2021*. Linköping University Electronic Press (LiU E-Press).

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

10. Language Resource References

Martin Malmsten and Love Börjeson and Chris Haffenden. 2020. *Swedish BERT models*. National Library of Sweden / KBLab. Distributed via HuggingFace.

Nivre, Joakim and Smith, Aaron. 2023. *Swedish-Talbanken-UD*. Universal Dependencies Consortium. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.13. PID http://hdl.handle.net/11234/1-5287.

Straka, Milan. 2023. *Universal Dependencies 2.12 models for UDPipe 2*. Distributed via LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University as part of Universal Dependencies 2.12. PID http://hdl.handle.net/11234/1-5200.

Volodina, Elena and Granstedt, Lena and Matsson, Arild and Megyesi, Beáta and Pilán , Ildikó and Prentice, Julia and Rosén, Dan and Rudebeck, Lisa and Schenström, Carl-Johan and Sundberg, Gunlög and Wirén, Mats. 2022. *SweLL-gold*. Språkbanken Text. Distributed via SBX/CLARIN. PID https://hdl.handle.net/10794/846.

The Vedic Compound Dataset

Sven Sellmer, Oliver Hellwig
Institute of Oriental Studies, Adam Mickiewicz University Poznań
Department of Comparative Language Science, University of Zürich
sven@amu.edu.pl, oliver.hellwig@uzh.ch

Abstract

This paper introduces the Vedic Compound Dataset (VCD), the first resource providing annotated compounds from Vedic Sanskrit, a South Asian Indo-European language used from ca. 1500 to 500 BCE. The VCD aims at facilitating the study of language change in early Indo-Iranian and offers comparative material for quantitative cross-linguistic research on compounds. The process of annotating Vedic compounds is complex as they contain five of the six basic types of compounds defined by Scalise and Bisetto (2005), which are, however, not consistently marked in morphosyntax, making their automatic classification a significant challenge. The paper details the process of collecting and preprocessing the relevant data, with a particular focus on the question of how to distinguish exocentric from endocentric usage. It further discusses experiments with a simple ML classifier that uses compound internal syntactic relations, outlines the composition of the dataset, and sketches directions for future research.

Keywords: Compounding, Sanskrit, Vedic, Dependency annotation

1. Introduction

Since the beginnings of modern linguistics, Sanskrit compounds have played a special role in research on compounding (see, e.g., Wujastyk, 1982), which is reflected by the fact that even some terms of the Indian grammatical tradition have entered current linguistic terminology (see Tab. 1). Sanskrit – especially its oldest form, known as Vedic, which was used from ca. 1500–500 BCE – is also of fundamental importance for Indo-European and cross-linguistic studies. Up until now, there exists a substantial collection of annotated compounds for classical and Neo-Sanskrit.[1] Many of these annotations, originating from works composed in the 19th and 20th c. CE, offer, however, only limited insights for historical linguistics due to their relatively recent composition. The Vedic Compound Dataset (VCD) introduced in this paper is the first resource to provide annotated compounds from Vedic, making it particularly well-suited for studying language change in the formative period of Sanskrit.[2]

2. Previous research

Quantitative cross-linguistic research on compounds has been less intensive than in other areas of linguistics (Moyna, 2019), but Guevara and Scalise (2009) have recently produced valuable statistics, in which, however, data for Indo-Aryan as well as ancient languages are lacking. The VCD fills both of these gaps to some extent and yields relevant comparative material (see Sec. 5).

Concerning annotation, Vedic compounds constitute an interesting challenge. As will be discussed in Section 4, they contain five of the six basic types of compounds that are used in Guevara and Scalise 2009. In addition, neither the compound internal relation between the words constituting them (see the examples in Section 3) nor the relation between a compound and the rest of the sentence are consistently marked in morphosyntax, which poses a challenge to their automatic classification. Over the past decade, several attemps at automatic classification of classical Sanskrit compounds have been undertaken. While Krishna et al. (2016) obtain 74% F-score for a dataset with four coarse compound categories by applying a Random Forest classifier to a set of manually defined linguistic markers, Sandhan et al. (2019) achieved a comparable F-score of 73% using an approach that combined a recurrent architecture with static word embeddings, bypassing the need for extensive feature engineering. Most recently, Sandhan et al. (2022) argued that compound classification needs to take syntactic properties of the surrounding text into account. They therefore combined compound classification with morphosyntactic tagging and dependency parsing in a joint learning task. Using a deep learning architecture with contextualized word embeddings, they report an F-score of 85.7% for coarse compound classification.

While these contributions have significantly advanced automatic Sanskrit compound classification, the present study did not use these systems for compound annotation for several reasons. Firstly, previous studies used classical Sanskrit data, but our focus is on Vedic compounds. The significant lexical differences between Vedic and classical Sanskrit can make applying these systems

[1] https://sanskrit.uohyd.ac.in/Corpus/

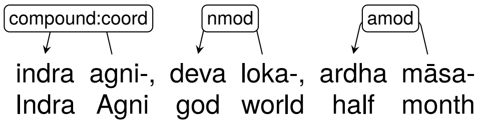
[2] The VCD is available at https://github.com/SvenSellmer/VedicCompoundDataset.

to Vedic texts problematic. Secondly, while an F-score of 85.7% is remarkable, it does not meet the high standards required for creating a reference dataset. Thirdly, the compound categories employed by these studies do not encompass all categories proposed by Bisetto and Scalise, limiting their applicability to our research. In what follows, we will present how we collected and prepared our data (Sec. 3 and 4), devoting particular attention to the recognition of their endocentric-exocentric dimension, and discuss experiments with a simple ML classifier. We will then discuss the composition of the dataset (Sec. 5) and draw conclusions for future research (Sec. 6).

3. Data collection

Our data is derived from two closely linked resources. The Digital Corpus of Sanskrit (Hellwig, 2010–2024) offers lexical and morphosyntactic annotations for Vedic and classical Sanskrit texts. Within the DCS, compounds that have a non-lexicalized reading (see below) are divided into their constituent parts. For instance, the coordinate compound *indrāgni-* ‘Indra and Agni’ is separated into the words *indra-* and *agni-*, each with its own morphosyntactic information. This preprocessing of the source data makes the identification of compounds significantly easier. The Vedic Treebank (VTB, Hellwig et al. 2020), containing approximately 32,000 sentences, supplements the DCS with a layer of Universal Dependencies (UD) annotations. The syntactic annotation of the VTB was carried out by a team of experts, who employed enhanced annotation guidelines (see Hellwig et al., 2023).

The standard UD guidelines offer only limited possibilities for a differentiated treatment of compounds,[3] which is unsatisfactory in view of the versatile role of compounds as an interface between syntax and lexicon and especially of the fact that Vedic compounds – like Sanskrit compounds in general (Lowe, 2015) – contain various syntactic structures, which tend to become diachronically increasingly complex. Therefore, the team extended the annotation guidelines (Biagetti et al., 2020) with the aim of enabling the annotator to make explicit the internal syntactic structure of a compound in the same way as UD labels show the relations obtaining between the words in a sentence. For instance, the compounds *indra-agni-*[4] ‘Indra and Agni’ (as a pair), *deva-loka-* ‘world of the gods’, and *ardha-māsa-* ‘half-month’ are annotated as follows:


compound:coord
nmod
amod
indra agni-, deva loka-, ardha māsa-
Indra Agni god world half month

The information – not immediately obvious in the latter two examples – that a word is a non-final member of a compound was incorporated into the VTB via the “Compound” feature (to be distinguished from the label `compound`, which is only used for coordinate compounds in the VTB). Compounds can include a limited number of particles and adverbs in addition to nominal forms (e.g. *sa-ratha-*, lit. ‘with-chariot’, i.e. “having a chariot”). Most of these indeclinables do not exist as standalone words. Since they constitute a closed lexical set, they can be directly integrated into compound detection. Adverbs that are part of compounds but do not belong to this closed set (e.g. *su-* ‘well’, which also occurs independently) were addressed individually during annotation.

To detect compounds in the VTB, we conducted a scan of the VTB’s conllu file for instances of the “Compound” feature and built compounds by tracing the syntactic arcs of the non-terminal members until we reached an inflected word form, which had to be the terminal member. In the example *deva-loka-* given above, *deva-* is labeled with the Compound feature in the VTB. Following the arc with the `nmod` label, we arrive at *loka-* which has an inflectional ending in a real world case and thus must constitute the terminal member of the compound.

The VCD is specifically designed to contain only two-word compounds. Apart from time restrictions, this focus is due to the fact that longer compounds of n words can typically be analyzed as multi-level mixed types consisting of $n-1$ elements. Furthermore, the oldest Vedic texts predominantly contain two-word compounds (see e.g. Macdonell, 1910, 143). By limiting our data selection to these short compounds, we ensure that our data covers the entire Vedic period. We equally did not include compounds that were identified as lexicalized by the annotator of the DCS. These compounds are typically technical terms. For instance, the term *agnihotra-* is a compound of the words *agni-* ‘sacrificial fire’ and *hotra-* ‘sacrificial libation’. However, an *agnihotra-* is not merely a ‘libation into the sacrificial fire’, but a specific type of such a libation (see e.g. Renou, 1953). Despite their semantic transparency, such lexicalized compounds are annotated as single words in the DCS and are not identified as compounds in its dictionary. As a result, we lack access to information indicating that *agnihotra-* is a compound, as well as its internal syntactic relation. The integration of such lexicalized compounds into the VCD remains an open issue for future research.

[3] See `https://universaldependencies.org/docs/en/dep/compound.html`.

[4] For convenience, all euphonic (‘sandhi’) changes have been removed in this paper.

	Endocentric	Exocentric
C.	Austria-Hungary (*dvandva*)	[lacking in E. and S.]
S.	horse-sacrifice (*tatpuruṣa*)	horse-faced (*bahuvrīhi*)
A.	blackbird (*karmadhāraya*)	redneck (*bahuvrīhi*)

Table 1: Compound classification according to Scalise and Bisetto 2005 (C. = coordinate; S. = subordinate; A. = attributive); indigenous terms in brackets.

Internal label	Compound type
`compound:coord`	→ coordinate
`nmod`, `obj`, `obl`, `iobj`	→ subordinate
`advmod`, `amod`, `nummod`, `acl`, `det`, `xcomp`, `nmod:appos`, `advcl`	→ attributive

Table 2: Internal labels and the dimension coordinate/subordinate/attributive.

4. Compound classification

Among the various possible classification schemes for compounds, we adopted the version proposed by Scalise and Bisetto (2005), for three reasons: 1. it is not only well-suited for Sanskrit (as can be seen in Biagetti, 2024) but also for many other languages; 2. it has already been employed in cross-linguistic studies (Scalise and Guevara, 2006; Guevara and Scalise, 2009), so that it facilitates the reusability of our dataset in such contexts; 3. it is convenient due to its conceptual simplicity, as opposed to its refined version in Scalise and Bisetto, 2009, which was too finegrained for our time budget.

This scheme has two-dimensions, as exemplified by the rows and columns of Tab. 1. Firstly it classifies compounds as endo- vs. exocentric, where an exocentric compound is understood as one that is not a hyponym of its formal head (see Bauer, 2017, 37, e.g., a redneck is not a kind of neck; for other definitions see Bauer, 2008 and Moyna, 2019). In Sanskrit grammar, these are called *bahuvrīhi*s: compounds that, as a whole, are (sometimes secondarily nominalized) adjectives though their final member is a noun. Secondly, it encodes the relation between the first and the final member, which may either be coordinate (i.e., dvandvas in the strict sense, Ralli 2019), subordinate, or attributive.

For the actual task of compound classification, the VCD provides the following information:

- UD label of the compound as a whole (i.e., of its final member)
- internal UD label
- POS information for both members
- case and gender of the final member

For classification according to this scheme we used an algorithm that was partly rule-based, and partly relied on human expertise:

1. The distinction between coordinate, subordinate, and attributive compounds could easily be made on the basis of the internal label, as shown in Table 2.

2. Coordinate compounds being endocentric by default in Sanskrit, subordinate und attributive compounds were then classified under the aspect of their exocentricity.

2a. In about 1/5 of the cases, this can be done automatically,[5] namely, where a mismatch between the gender of the compound and the gender of its final member as an independent noun can be observed. For instance, the compound *aśva-mukha-* can be either endocentric ('face of a horse') or exocentric ('horse-faced'). Now, *mukha-* 'face' is a neutral noun, so wherever *aśva-mukha-* features a non-neuter ending it must refer as an adjective to a masculine or feminine noun (e.g., *aśva-mukhaḥ rākṣasaḥ* 'a horse-faced demon') and so be an exocentric compound.

2b. Further, a sizeable subgroup of exocentric compounds (ca. 600 tokens) could be classified on the basis of their morphology: the so-called root or synthetic compounds with a verbal root noun as final member are always exocentric (Scarlata, 1999); e.g., *prathama-ja* 'first-born', from $\sqrt{jan}$ 'to be born'. Detecting them could not be fully automatized as there is no appropriate POS tag in the DCS flagging them as verbal roots.

2c. In the remaining ca. 1,700 cases, the decision to classify a given compound as exo- or endocentric could only be made by a human expert on the basis of its use in the actual context. Dictionary information could be used in cases in which the translation indicated exclusively exo- or endocentric usage. But such hints were not available for all compounds, and in addition turned out to be not always reliable. Opposite to what one may expect, the UD label of the final compound member did not allow to decide between exo- and endocentric usage. For example, in their prototypical role as adnominal modifiers, *bahuvrīhi*s are linked by `acl` to their referents (Biagetti et al., 2020, Sec.

[5] In accented texts, also the location of the accent in a compound often is indicative of exocentricity (Wackernagel, 1905, § 113), but this information was not available to us as it is lacking in the DCS.

2.7.2). However, even this label cannot serve as a reliable indicator of exocentricity, because it also appears with endocentric compounds, for instance, in relative clauses. In addition, only about 30% of all *bahuvrīhi*s are used as adnominal modifiers, as they are, for instance, often substantivized and function as independent nouns. As a consequence, the annotation of these 1,700 compounds had to be done manually, which turned the present step into the most time-consuming one.

The description of the annotation process suggests that many decisions are rule-based, i.e., can be made based on the internal and external syntactic relations of compounds and their morphosyntactic information. We hypothesized that a simple classification algorithm with access to the syntactic gold information of the VCD could learn these rules. To test this hypothesis, we implemented a multinomial regression model. The predictors for this model include the aforementioned compound-internal and external syntactic labels, as well as the part-of-speech tags and lemmata of the two words constituting a compound. The model is trained to predict which of the five classes in the scheme of Bisetto and Scalise (Table 1) a compound belongs to. The results of a tenfold cross-validation (see Table 3) show that the system achieves F-scores above 80%, even for complicated classes that involve decisions between endo- and exocentric use. As the F-scores of the two ablation tests in columns 5 and 6 of Table 3 (-I: no compound internal syntactic labels; -O: no outer labels) indicate, this success is mainly due to the availability of compound-internal syntactic relations from the VTB. While ignoring the labels that connect compounds with the rest of the sentence and thus indicate their syntactic function (-O) keeps the F-scores largely unchanged, ignoring their inner syntactic labels (-I) leads to substantially lower F-scores for three out of five types. Specifically, the low $\text{F}_{\text{-I}}$-score for coordinate endocentric compounds likely results from the fact that they are not distinguished by POS information from subordinate compounds, but occur with much lower frequency (see Tab. 5.) These findings can inform future research in automatic compound classification.

Type	P_{All}	R_{All}	F_{All}	$\text{F}_{\text{-I}}$	$\text{F}_{\text{-O}}$
attrib/endo	81.8	86.4	84.0	80.2	79.9
attrib/exo	81.0	80.9	80.9	74.8	80.0
coord/endo	97.6	98.6	98.1	29.6	98.1
subord/endo	91.4	92.3	91.8	82.3	90.5
subord/exo	87.2	81.1	84.1	80.0	78.8

Table 3: Results of the multinomial classifier for compound types, 10-fold cross-validation. All: all predictors, -I: no internal syntactic labels, -O: no outer labels.

Deprel	#Tok.	Deprel	#Tok.
`nmod`	2260	`nummod`	574
`advmod`	1089	`obl`	460
`amod`	800	`acl`	191
`obj`	721	`det`	189
`compound:coord`	632	`iobj`	26

Table 4: Most frequent compound-internal dependency relations in the VCD.

5. Composition of the dataset

The VCD contains almost 7,000 two-word compounds together with information on morphology, internal and external syntactic relations, chronology, and Vedic subtraditions. A few plots and tables may serve to give an overview of the composition of our dataset. Tab. 4 lists the most frequent compound internal labels in the VCD. It thus gives insights into the syntactic processes active during compounding and so can serve as a starting point for cross-linguistic comparison and for the construction of fine-grained semantic frames. Tab. 5 shows the distribution of the tokens over Scalise and Bisetto's classification. The numbers confirm the general cross-linguistic observations in Guevara and Scalise 2009, 118–119, that in terms of frequency S. > A. > C. Regarding the endo-/exocentric distinction, exocentric compounds make up 41.8% of all compounds in the VCD. This is a remarkably high percentage compared with the statistics in Scalise et al. 2009, where this ratio ranges from 8.4% (Germanic languages) to 35.4% (Romance languages). The ratio for Vedic gets even higher when the diachronic dimension of our dataset is taken into account. As can be seen in Fig. 1, right, it drops from an extreme ratio of 72.4% in the archaic Rig Vedic period, a figure reminiscent of what Bauer (2008, 68) reports for some African and Australian languages, to 30.3% in the late Sūtras. Notably, this development runs counter to the general rise in compound usage, as shown in Fig. 1, left.

	Endocentric		Exocentric		All	
	Tok.	%	Tok.	%	Tok.	%
C.	632	9.0	0	0	632	9.0
S.	2,273	32.5	1,177	16.8	3,450	49.3
A.	1,166	16.7	1,744	24.9	2,910	41.6
	4,071	58.2	2,921	41.8	6,992	

Table 5: Counts of the main compound categories (tokens) in the VCB.

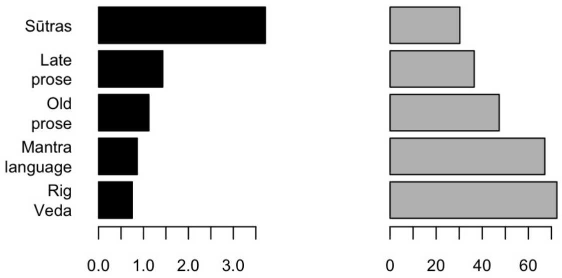


Figure 1: Percentages of compounds among all lemmata (= left) and of exocentric compounds among all compounds (= right), across Vedic literary periods; earliest in the bottom-most row.

6. Conclusion

Up until now, the diachronic, geographical and sociolinguistic development of Vedic literature remains incompletely understood (Witzel, 1997). The compounds collected in the VCD, showing clear diachronic trends regarding their endo-/exocentric dimension (see Fig. 1), thus provide valuable data for gaining deeper insights into the linguistic developments during this period as well as for timestamping Vedic texts (Hellwig, 2024). They can further prove fruitful for comparative studies in an Indo-European and cross-linguistic framework, as they contain data about one of the earliest attested Indo-European languages.

The rule-based parts of collecting the dataset were comparatively straightforward, but to distinguish between exocentric and endocentric compounds of the attributive and subordinate types turned out to be a time-consuming process. It is important to note that this work would have been unnecessary if such a distinction could be directly established on the basis of the UD labels. It would be therefore helpful to add an appropriate UD sublabel to, e.g., `nmod` and `amod`, to indicate *bahuvrīhi*s in various languages. This would be a small extra effort, because for a human expert annotating a whole sentence it is usually evident whether a given compound is exocentric. It is to be expected that DL dependence parsers will then be able to process these annotations and to determine the exocentricity of compounds with high precision. This would be a highly desirable outcome for the research on compounds in general, as their exocentric/endocentric dimension is of fundamental importance. In addition, due to the general tendency of exocentric compounds for having a metonymic meaning (Bauer, 2008; Barcelona, 2008), such a sublabel would also be relevant for metonymy recognition.

7. Ethical considerations

We are not aware of any ethical issues arising from the composition or use of our data set.

8. Limitations

Four limitations of our dataset should be mentioned. Firstly, it must be understood that – though of considerable size for an ancient language – it is based on only about 35% of the extant Vedic literature – nevertheless, its chronologically balanced composition and the wide variety of texts it draws on make it useful for quantitative linguistic studies. Secondly, as discussed on p. 2 above, we did not consider compounds that were treated as lexicalized in the DCS. Thirdly, for the reasons explained on p. 2, we restricted ourselves to two-word compounds for the time being. We plan to overcome these limitations in future versions of the VCD. Finally, it should be noted that the POS tags taken over from the VTB are not completely reliable. In the VCD, they have been manually corrected in a number of instances, but not in the form of a systematic revision.

9. Acknowledgements

We would like to express our gratitude to three anonymous reviewers, who made a number of very helpful remarks and suggestions.

Research for this paper was funded by the German Federal Ministry of Education and Research, FKZ 01UG2121.

10. Bibliographical References

Antonio Barcelona. 2008. The interaction of metonomy and metaphor in the meaning and form of ‘bahuvrīhi’ compounds. *Annual Review of Cognitive Linguistics*, 6:208–281.

Laurie Bauer. 2008. Exocentric compounds. *Morphology*, 18(1):51–74.

Laurie Bauer. 2017. *Compounds and Compounding*. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge.

Erica Biagetti. 2024. Early Vedic compounds: A typological reappraisal. *Studies in Language*, 48(1):1–64.

Erica Biagetti, Oliver Hellwig, Salvatore Scarlata, Paul Widmer, and Elia Ackermann. 2020. Annotation guidelines for the Vedic Treebank. v2.0 – July 2020.

Emiliano Guevara and Sergio Scalise. 2009. Searching for universals in compounding. In Sergio Scalise, Elisabetta Magni, and Antonietta Bisetto, editors, *Universals of language today*, pages 101–128. Springer, Amsterdam.

Oliver Hellwig. 2010–2024. DCS - The Digital Corpus of Sanskrit.

Oliver Hellwig. 2024. To compound or not to compound? A diachronic Bayesian analysis of compounds and equivalent constructions in Vedic Sanskrit. *Indogermanische Forschungen*, 129.

Oliver Hellwig, Sebastian Nehrdich, and Sven Sellmer. 2023. Data-driven dependency parsing of Vedic Sanskrit. *Language Resources and Evaluation*, 57:1173–1206.

Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. The treebank of Vedic Sanskrit. In *Proceedings of the LREC*, pages 5139–5148.

Amrith Krishna, Pavankumar Satuluri, Shubham Sharma, Apurv Kumar, and Pawan Goyal. 2016. Compound type identification in Sanskrit: What roles do the corpus and grammar play? In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 1–10, Osaka, Japan. The COLING 2016 Organizing Committee.

John J. Lowe. 2015. The syntax of Sanskrit compounds. *Language*, 91(3):71–115.

Arthur Anthony Macdonell. 1910. *Vedic Grammar*. Trübner, Strassburg.

María Irene Moyna. 2019. Exocentricity in morphology. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Angela Ralli. 2019. Coordination in compounds. In *Oxford Research Encyclopedias – Linguistics*. Oxford University Press.

Louis Renou. 1953. *Vocabulaire du rituel védique*. Librairie C. Klincksieck, Paris.

Jivnesh Sandhan, Ashish Gupta, Hrishikesh Terdalkar, Tushar Sandhan, Suvendu Samanta, Laxmidhar Behera, and Pawan Goyal. 2022. A novel multi-task learning approach for context-sensitive compound type identification in Sanskrit. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4071–4083, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jivnesh Sandhan, Amrith Krishna, Pawan Goyal, and Laxmidhar Behera. 2019. Revisiting the role of feature engineering for compound type identification in Sanskrit. In *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 28–44, IIT Kharagpur, India. Association for Computational Linguistics.

Sergio Scalise and Antonietta Bisetto. 2005. The classification of compounds. *Lingue e Linguaggio*, 2:319–332.

Sergio Scalise and Antonietta Bisetto. 2009. The classification of compounds. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford handbook of compounding*, Oxford handbooks in linguistics, pages 34–53. Oxford University Press, Oxford ; New York.

Sergio Scalise, A. Fábregas, and F. Forza. 2009. Exocentricity in compounding. *Gengo Kenkyu*, 135:49–84.

Sergio Scalise and Emiliano Guevara. 2006. Exocentric compounding in a typological framework. *Lingue e linguaggio*, 2:185–206.

Salvatore Scarlata. 1999. *Die Wurzelkomposita im Ṛg-Veda*. Reichert, Wiesbaden.

Jakob Wackernagel. 1905. *Altindische Grammatik. Band II, 1: Einleitung zur Wortlehre. Nominalkomposition*. Vandenhoeck & Ruprecht, Göttingen.

Michael Witzel. 1997. The development of the Vedic canon and its schools: The social and political milieu. In Michael Witzel, editor, *Inside the Texts, Beyond the Texts*, pages 257–345. Department of Sanskrit and Indian Studies, Harvard University, Cambridge, Mass.

Dominik Wujastyk. 1982. Bloomfield and the Sanskrit origin of the terms ‘exocentric’ and ‘endocentric’. *Historiographia Linguistica*, 9(1):179–184.

A Universal Dependencies Treebank for Gujarati

Mayank Jobanputra[1*], **Maitrey Mehta**[2*], **Çağrı Çöltekin**[3]
[1]Department of Language Science and Technology, Saarland University
[2]Kahlert School of Computing, University of Utah
[3]Department of Linguistics, University of Tübingen
mayank@lst.uni-saarland.de, maitrey@cs.utah.edu, ccoltekin@sfs.uni-tuebingen.de

Abstract

The Universal Dependencies (UD) project has presented itself as a valuable platform to develop various resources for the languages of the world. We present and release a sample treebank for the Indo-Aryan language of Gujarati – a widely spoken language with little linguistic resources. This treebank is the first labeled dataset for dependency parsing in the language and the script (the Gujarati script). The treebank contains 187 part-of-speech and dependency annotated sentences from diverse genres. We discuss various idiosyncratic examples, annotation choices and present an elaborate corpus along with agreement statistics. We see this work as a valuable resource and a stepping stone for research in Gujarati Computational Linguistics.

Keywords: low-resource languages, universal dependencies, Gujarati

1. Introduction

The Universal Dependencies (UD) project (Nivre et al., 2016; de Marneffe et al., 2021) offers cross-linguistically consistent annotations for dependency treebanks, part-of-speech, and morphological features. The ever-expanding language base under the UD umbrella ensures that similar language patterns can be dealt with consistently when working with a new language. Further, language-specific features are brought to the fore for discussion. As a result, UD becomes the most fundamental of resources to be developed for a particular language.

Gujarati is an Indo-Aryan language originating from the western Indian state of Gujarat. The language is widely spoken by over 56 million speakers (Eberhard et al., 2022) and is one of the 22 languages with official status in India. Yet, the Gujarati Computational Linguistics community is still in its infancy. Joshi et al. (2020) classify Gujarati in the “Scraping-Bys” category (category 1) in their taxonomy indicating a scant availability of labeled datasets. Basic resources such as part-of-speech taggers, and named entity recognizers are not readily available. Hence, a dependency treebank in such a language can have a wide-reaching impact.

On the other hand, the UD community has already produced a handful of treebanks in various Indo-Aryan languages. As a result, we are equipped with resources in related languages like Marathi (Ravishankar, 2017), Hindi (Bhat et al., 2017; Zeman et al., 2017), and Punjabi (Arora, 2022). Such resources are of value while constructing a sample Gujarati treebank.

The benefits of building a sample Gujarati treebank are four-fold:

a) It presents as a valuable resource for the development of linguistic tools and resources in a low-resource language, i.e., Gujarati.

b) Gujarati uses a unique eponymous script that is not yet represented in the UD project. This can be especially valuable for future researchers interested in building resources for lesser-resourced languages such as Kutchi, and Bhili that also use the Gujarati script.[1]

c) It ensures annotation paradigms in similar contexts are adhered to and helps point out any discrepancies in existing treebanks.

d) We can point out some new idiosyncratic phenomena that might be Gujarati-specific, or missed by earlier works.

The above-mentioned reasons motivate us to propose a sample dependency treebank for Gujarati: *GujTB*.[2] In the subsequent sections, we explain the selected corpora, statistics and highlight some interesting discussion points encountered.

2. The Dataset

In this section, we provide details of the annotated corpora and the annotation process.

Corpora. We investigated available corpora that include Gujarati text such as IndicCorp (Kakwani et al., 2020) and Samanantar (Ramesh et al., 2022). However, we observe that these datasets majorly contain news and other formal

[*] Both authors contributed equally.

[1] https://www.omniglot.com/writing/languages.htm

[2] Code & Data available at: https://github.com/UniversalDependencies/UD_Gujarati-GujTB

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 56–62. May 25, 2024.

texts. Hence, we annotate a total of 187 sentences taken from diverse sources like Samanantar (news), UD Cairo (short),[3] Gujarati translations (from Mehta and Srikumar, 2023) of the French novella – *Le Petit Prince* (fiction) (The Little Prince, de Saint-Exupéry, 1943), and a Gujarati grammar book (grammar) (Raimond, 2004).

Annotation Process and Agreement. Two of the paper authors[4] annotated this dataset. The annotations were created separately, and followed by an initial correction phase to fix any obvious errors. A hundred-sentence subset of annotations was considered for the inter-annotator agreement (IAA) study.[5] The IAA for the part-of-speech (POS) tags is 99.87 (Cohen's κ). The head selection agreement is 99.44% and the relation agreement on the heads that matched is 99.88 (Cohen's κ). The head selection agreement is the proportion of dependents assigned the same head by both annotators (similar to the unlabeled attachment score).

Dataset Statistics. The dataset statistics by genre are given in Table 1. The distribution of POS tags in the corpus is given in Table 2. Furthermore, we provide the statistics regarding dependency relations in Table 3. Notably, our dataset is a representative set of all possible relations in Gujarati.

Genre	Sentences	Tokens
news	93	1159
short	20	178
fiction	40	331
grammar	34	217
Total	187	1885

Table 1: Data statistics by genre for GujTB.

POS	Counts
NOUN	425
PUNCT	250
VERB	213
AUX	185
ADP	152
PROPN	145
ADJ	134
PRON	133
ADV	60
CCONJ	50
PART	43
NUM	40
DET	23
INTJ	14
SCONJ	13
SYM	3
X	2
Total	1885

Table 2: Part-of-speech tag statistics.

Relation	Counts
punct	250
root	187
nsubj	174
case	151
aux	133
nmod	129
obl	110
obj	99
amod	96
compound	70
advmod	62
conj	59
cc	52
cop	51
discourse	44
flat	36
advcl	35
nummod	27
det	21
acl	17
mark	14
ccomp	13
appos	13
parataxis	13
iobj	11
orphan	3
dislocated	3
goeswith	3
fixed	2
xcomp	2
vocative	1
reparandum	1
Total	1885

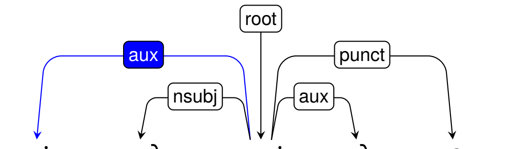
Table 3: Dependency relation statistics. All relation sub-types have been merged with their universal classes for representation.

3. Syntactic Relations

In this section, we discuss the many interesting dependency choices. While a large volume of dependency choices such as subjects, object, and light/serial verb constructions follow existing Indo-Aryan literature (Bhat et al., 2017; Ravishankar, 2017; Ojha and Zeman, 2020; Arora, 2022), our goal is to highlight the more subjective cases.

Interrogative/Question particles. The treatment of interrogative or question particles has largely varied in the UD literature.[6] We follow the preceding Indo-Aryan treebanks in assigning question particles with the respective dependency and POS tags as what would be assigned for a valid answer substitution. However, in cases where an obvious substitution is not viable (e.g., Yes/No questions) as shown in Example 1, we find that an aux relation fits the best.

(1)



શું	તારે	જવું	છે	?
shuṃ	*tare*	*javuṃ*	*che*	*?*
Do	*you.ERG*	*go.DES*	*is*	*?*
AUX	*PRON*	*VERB*	*AUX*	*PUNCT*

'Do you want to go ?'

[3] https://github.com/UniversalDependencies/cairo

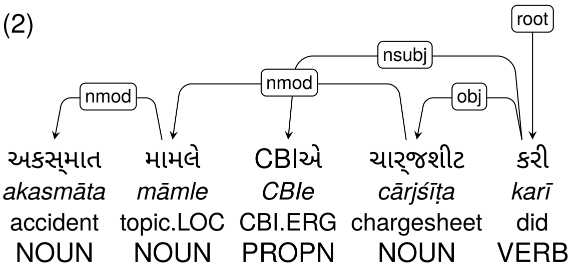
[4] Both are L1 speakers of Gujarati

[5] We release both the individual and adjudicated dataset as per Plank (2022)'s suggestion.

[6] https://github.com/UniversalDependencies/docs/issues/738

Non-projectivty. Bhat et al. (2017, pp.23) discuss non-projectivity in Hindi. Gujarati allows non-projective trees in a similar spirit. Partial free word order as shown in Example 2 can give rise to overlapping dependency edges.

(2)

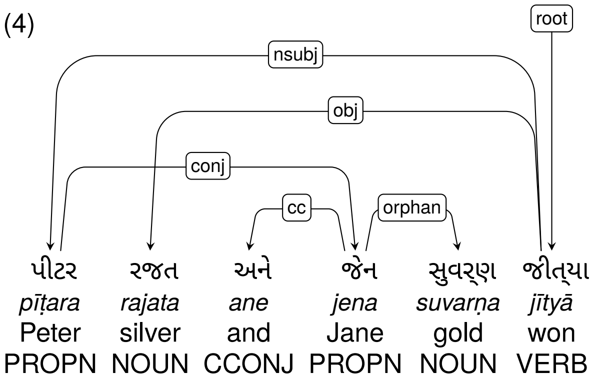

root
nsubj
nmod
nmod
obj
અકસ્માત akasmāta accident NOUN
મામલે māmle topic.LOC NOUN
CBIએ CBIe CBI.ERG PROPN
ચાર્જશીટ cārjśīṭa chargesheet NOUN
કરી karī did VERB

‘CBI made a chargesheet about the accident’

Head-final conjunctions. UD guidelines necessitate that the head of a conjunctive phrase be the first conjunct. However, Gujarati carries case inflections and post-positional attachments on the final conjunct which mediate semantic relations between the governor and the conjunctive phrase (see Example 3). This may lead to unwarranted non-projectivity as shown in Example 4.

Note that, in Example 4, the English translation fails to mark plurality on the verb “*won*” while in Gujarati “*jītyā*” has a plural inflection. As a result, the entire conjunctive phrase, not individual proper nouns (*Peter* or *Mary*), has to be the subject. At first sight, the non-projectivity in this example may seem avoidable by annotating promoted subject “*pīṭara*” as `root`, and attaching “*rajata*” to “*pīṭara*” as `orphan`, with the second clause attached as `conj` to the first clause. However, this would cause the plural verb to agree with a singular subject which is not the head of the coordinated structure. Similar issues also arise due to fixed head-initial coordination rule in UD for other head-final languages (Çöltekin, 2015; Kanayama et al., 2018; Tyers et al., 2017; Han et al., 2020). Hence, an argument can be made to mark the final conjunct as the head of the conjunctive phrase. However, we follow the UD guidelines and mark the first conjunct to be the head of the phrase.

(4)

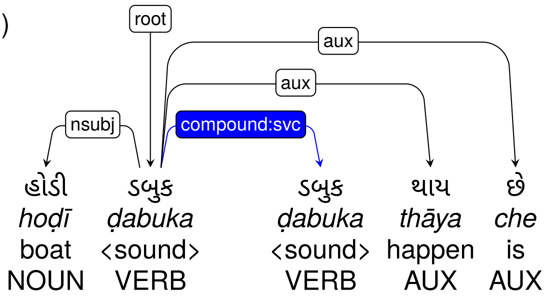


‘Peter won silver and Jane gold’

Polarity/emphatic markers within serial verb constructions. Gujarati supports verb-verb constructions where the second verb is, usually, semantically bleached. Owing to the existence of partial free-word ordering discussed before, we observe that serial verb constructions are often separated by polarity or emphatic particles as seen in Example 5. To the best of our knowledge, this case is idiosyncratic to Gujarati. However, note that the treatment of these particles does not change.

Ideophonic verbs. In Gujarati, repetitions of a word can occur in two cases: discursive repetitions (બોલ બોલ [“tell tell”], જા જા [“go go”]) and onomatopoeias (ધમ ધમ [“dham dham”], the sound of Indian drums). Example 6 presents a case of onomatopoeias. Szubert et al. (2021) introduced `parataxis:repeat` for expressing adjectival repetitions in child-directed speech. Sulubacak et al. (2016) use `compund:redup` for reduplicated words. In our case, onomatopoeias are used to imitate different sounds that express actions and act as verbal repetitions. Hence, we suggest using `compound:svc`. To indicate the ideophonic nature of the verb, we mark the feature `VerbType=Ideo`.[7]

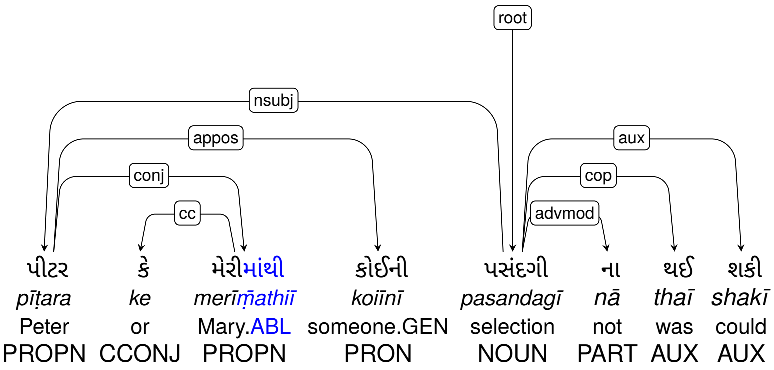
(6)



‘The boat is bobbling’

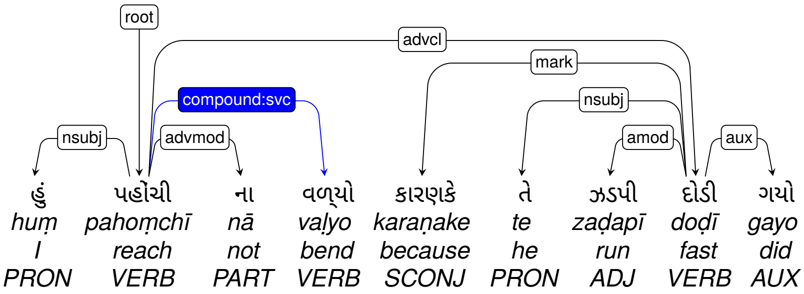
[7] As noted in https://github.com/UniversalDependencies/docs/issues/842

(3)


root
nsubj
appos
aux
conj
cop
cc
advmod
પીટર કે મેરીમાંથી કોઈની પસંદગી ના થઈ શકી
pīṭara ke merīṃathiī koiīnī pasandagī nā thaī shakī
Peter or Mary.ABL someone.GEN selection not was could
PROPN CCONJ PROPN PRON NOUN PART AUX AUX

‘No one from Peter or Mary got selected’

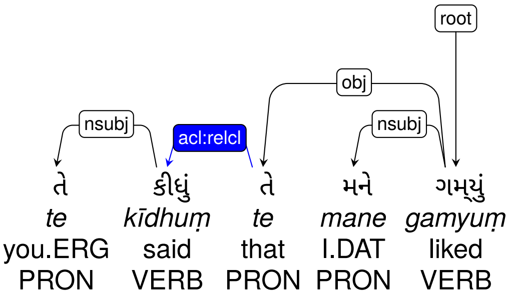
(5)


root
advcl
mark
compound:svc
nsubj
nsubj
advmod
amod
aux
હું પહોંચી ના વળ્યો કારણકે તે ઝડપી દોડી ગયો
huṃ pahoṃchī nā vaḷyo karaṇake te zaḍapī doḍī gayo
I reach not bend because he run fast did
PRON VERB PART VERB SCONJ PRON ADJ VERB AUX

‘I could not keep up because he ran fast’

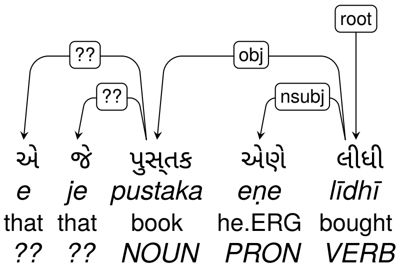
Absence of clausal subjects. We find that clausal subjects do not exist in Gujarati. We substantiate this argument using an English example, *“What she said is likable.”*: i) A perfect translation of this sentence does not exist in Gujarati. A close translation is given in Example 7. Note that a co-referential pronominal તે [te, that] is added to construct a grammatically sound sentence. ii) Secondly, the presence of a dative nominal construction with experiencer semantics is permitted. Such constructions are considered grammatical subjects (Arora, 2022) which makes clausal subjects impossible. iii) Finally, the mandatory co-referential pronominal mediates the relation between the governor and the would-be subject clause.

(7)



‘What you said is liked by me’

(8)



‘That book which he took’

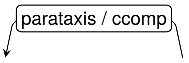
Challenging Construction. Example 8 depicts a case where arguments can be made for multiple possible annotations: i) Assigning `det:predet` to એ [e] and `det` જે [je] with પુસ્તક [pustaka] as their head ii) One may argue a change in order between “જે” and “પુસ્તક”, where “જે” would act as a subordinating conjunction. However, we contend a semantic difference between this sentence and the one presented in Example 8. We lean towards the first annotation.

Quoter and Quotation. We encounter a screenplay dialog-style quotation that is yet to be resolved (see Example 9).[8] Recent guidelines recommend `ccomp` over `parataxis` for reported

[8] This is not a Gujarati-specific issue. Moreover, we have opened a discussion regarding this point:

speech.[9] We believe this to be a much more pervasive (and not a Gujarati-specific) issue; applicable, perhaps, when UD is extended to plays.

(9)



I play football : Mark
'I play football : Mark'

4. Tokenization and Part of Speech

Splitting Genitive Markers. Certain nominals (and, in some instances, verbs) in Gujarati are inflected for case. It is unclear if these suffixes should be separated from their heads. This is a known issue that has been raised in Ravishankar (2017). They choose to split genitive markers to be consistent with Hindi. We follow the same rule with the added incentive to separate out layer III postpositions that pair postpositions with preceding genitive markers (Masica, 1993).

The Case for Determiners. According to Gujarati grammars (Tisdall, 1892; Doctor, 2004), demonstrative pronouns like એ [e], તે [te], પેલું [peluṃ], etc. behave differently when attached to a nominal, versus when used independently. When occurring independently, we treat them as pronouns. Tisdall (1892) argues to treat them as adjectives when used with nominals (e.g., એ કૂતરો 'that dog'). Gujarati grammar does not discuss determiners as such. However, we see this usage closer to the UD definition of determiners and hence use the same.

Modal auxiliaries. There are several verbs that can be compounded with other verbs, nouns, or adjectives to form verb compounds. While most of these are semantically bleached, Gujarati identifies a fixed set of verbs to act as modal auxiliaries (Doctor, 2004). This fixed set includes verbs like 'જા [jā,go], આવ [āva,come], રહે [rahe,stay]' (temporal), 'કર [kara,do], લાગ [lāga,feel]' (compulsion), and 'પડ [pada,fell],જોઈ [joī,want]' (obligation). We mark these fixed set of verbs as auxiliaries while the rest are marked as regular verbs.

5. Conclusion and Future Work

We present the first dependency treebank in the Gujarati language and script. We provided detailed dataset statistics and discussed interesting examples and decisions. In a low-resourced language like Gujarati, we see this sample treebank as an enabler for future computational linguistics research. In the future, we aim to increase the size of the annotated corpora to help contribute a dependency parser. Furthermore, we also intend to provide annotations for the morphological features of Gujarati.

6. Ethics Statement

The dataset presented in this work is a voluntary annotation effort between the two authors of this paper. While the annotators speak different dialects of Gujarati, we are aware that our corpus might not contain diverse dialectical varieties.

Acknowledgements

We thank Dr. Atul Kr. Ojha and Aryaman Arora for their useful insights. We thank the anonymous reviewers for their valuable feedback. We highly appreciate the feedback received at the UniDive (CA21167) General Meeting 2023 where we presented an extended abstract version of this work. The format of the paper is inspired by Arora (2022).

7. Bibliographical References

Aryaman Arora. 2022. Universal Dependencies for Punjabi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5705–5711, Marseille, France. European Language Resources Association.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia. 2017. *The Hindi/Urdu Treebank Project*, pages 659–697. Springer Netherlands, Dordrecht.

Çağrı Çöltekin. 2015. A grammar-book treebank of Turkish. In *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 35–49.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

de Saint-Exupéry, Antoine. 1943. *Le petit prince [The little prince]*. Reynal & Hitchkock (US), Gallimard (FR).

Doctor, Raimond. 2004. *A Grammar of Gujarati*. Lincom Europa.

https://github.com/UniversalDependencies/docs/issues/904

[9] https://universaldependencies.org/changes.html#reported-speech

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World. Twenty-fifth edition.* SIL International, Dallas, Texas.

Ji Yoon Han, Tae Hwan Oh, Lee Jin, and Hansaem Kim. 2020. Annotation issues in Universal Dependencies for Korean and Japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 99–108, Barcelona, Spain (Online). Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Kakwani, Divyanshu and Kunchukuttan, Anoop and Golla, Satish and N.C., Gokul and Bhattacharyya, Avik and Khapra, Mitesh M. and Kumar, Pratyush. 2020. *IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages*. Association for Computational Linguistics.

Hiroshi Kanayama, Na-Rae Han, Masayuki Asahara, Jena D. Hwang, Yusuke Miyao, Jinho D. Choi, and Yuji Matsumoto. 2018. Coordinate structures in Universal Dependencies for head-final languages. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 75–84, Brussels, Belgium. Association for Computational Linguistics.

Colin P Masica. 1993. *The indo-aryan languages*. Cambridge University Press.

Mehta, Maitrey and Srikumar, Vivek. 2023. *Verifying Annotation Agreement without Multiple Experts: A Case Study with Gujarati SNACS*. Association for Computational Linguistics.

Luís Morgado da Costa, Francis Bond, and Roger V. P. Winder. 2022. The Tembusu Treebank: An English Learner Treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4817–4826, Marseille, France. European Language Resources Association.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Atul Kr. Ojha and Daniel Zeman. 2020. Universal Dependency Treebanks for Low-Resource Indian Languages: The Case of Bhojpuri. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 33–38, Marseille, France. European Language Resources Association (ELRA).

Barbara Plank. 2022. The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Doctor Raimond. 2004. A grammar of gujarati. *München: Lincom Europa. Search in*.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Vinit Ravishankar. 2017. A Universal Dependencies treebank for Marathi. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 190–200, Prague, Czech Republic.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan. The COLING 2016 Organizing Committee.

Ida Szubert, Omri Abend, Nathan Schneider, Samuel Gibbon, Sharon Goldwater, and Mark Steedman. 2021. Cross-linguistically Consistent Semantic and Syntactic Annotation of Child-directed Speech. *arXiv preprint arXiv:2109.10952*.

Tisdall, WS. 1892. *A Simplified Grammar of the Gujarati Language*. Sagwan Press.

Francis Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An assessment of Universal Dependency annotation guidelines for Turkic languages. In *5th International Conference on Turkic Language Processing (TURKLANG 2017)*, pages 356–377.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Overcoming Early Saturation on Low-Resource Languages in Multilingual Dependency Parsing

Jiannan Mao[†,*], **Chenchen Ding**[‡], **Hour Kaing**[‡],
Hideki Tanaka[‡], **Masao Utiyama**[‡], **Tadahiro Matsumoto**[†]
[†]Gifu University, Gifu, Japan
[‡]National Institute of Information and Communications Technology, Kyoto, Japan
[†]{mao, tad}@mat.info.gifu-u.ac.jp
[‡]{chenchen.ding, hour_kaing, hideki.tanaka, mutiyama}@nict.go.jp

Abstract

UDify (Kondratyuk and Straka, 2019) is a multilingual and multi-task parser fine-tuned on mBERT that achieves remarkable performance in high-resource languages. However, the performance saturates early and decreases gradually in low-resource languages as training proceeds. This work applies a data augmentation method and conducts experiments on seven few-shot and four zero-shot languages. The unlabeled attachment scores were improved on the zero-shot languages dependency parsing tasks, with the average score rising from 67.1% to 68.7%. Meanwhile, dependency parsing tasks for high-resource languages and other tasks were hardly affected. Experimental results indicate the data augmentation method is effective for low-resource languages in a multilingual dependency parsing.

Keywords: Parsing, Multilinguality, Low Resource Languages, Unsupervised Learning

1. Introduction

A dependency parser can be efficiently trained when large treebanks are available (Dozat and Manning, 2017; Qi et al., 2020). For low-resource languages with no (zero-shot) or limited (few-shot) treebanks, multilingual modeling has emerged as an efficient solution, where cross-lingual information is leveraged to compensate for the lack of data. Scholivet et al. (2019); Üstün et al. (2022) have demonstrated that the performance on multilingual tasks can be boosted by pairing languages with similarities. Multilingualism also reduces the expense when training multiple models for a group of languages (Johnson et al., 2017; Aharoni et al., 2019; Cai et al., 2021; Muennighoff et al., 2023).

UDify (Kondratyuk and Straka, 2019) is a multitask network fine-tuned on multilingual BERT (mBERT) (Devlin et al., 2019) pre-trained embeddings. It is capable of producing annotations for any treebank from Universal Dependencies (UD) (Zeman et al., 2018). UDify exhibits strong and consistent performance across all 124 UD treebanks for 75 languages and multiple tasks such as lemmatization, part-of-speech (POS), and dependency parsing. However, an issue not yet paid enough attention in several related studies is the substantial discrepancy found in the performance of these methods in low-resource language learning scenarios, even when almost the identical training strategies, datasets, models, and evaluation methods were used in Choudhary (2021), Üstün et al. (2022), Effland and Collins (2023).

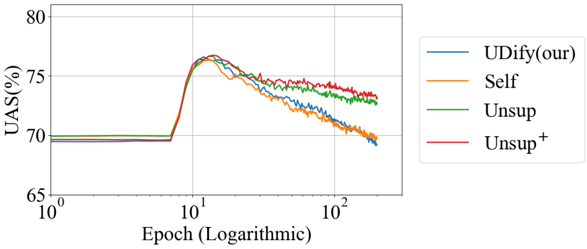


Figure 1: Change in the UAS(%) of a model during the training process on the Breton–KEB test set for both baselines: UDify(our) and Self, as well as the proposed method, Unsup and Unsup$^+$.

To address and investigate this issue, the work of Mao et al. (2023) conducts an experimental exploration into the low-resource case phenomenon by observing changes during model training. They adopted the data augmentation strategy, which leverages the original UDify for parsing raw sentences in single low-resource language to obtain initial probabilities. This is followed by the application of unsupervised learning to train these probabilities. Using the trained probabilities to create artificially structured dependency data and merging them into UDify's training set enables UDify to be trained on a more extensive dataset.

In this work, we conducted comprehensive experiments on low-resource languages using data augmentation methods, expanded (for few-shot languages) and created (for zero-shot languages) artificial treebanks for the seven few-shot and four

[*] This work was done during the first author's internship at National Institute of Information and Communications Technology, Kyoto, Japan.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 63–69. May 25, 2024.

zero-shot languages. By combining these artificial treebanks with the UD treebanks and using the UDify framework, we trained a multilingual parser. As a result, increases in the unlabeled attachment score (UAS) for zero-shot languages were observed, with the average value increasing from 67.1% to 68.7%; in the most-improved case, the UAS rocketed from 78.4% to 88.0%. Similarly, the few-shot languages experienced a UAS increase of 0.2%. In contrast, the UAS for other languages and evaluation scores for other tasks did not show significant changes, which suggests that the overall robustness of multilingual and multi-task processing is retained.

2. Background

2.1. UDify

The UDify model jointly predicts lemmas, POS tags, morphological features, and dependency structures. The pre-trained mBERT model[1] is used in the UDify model for cross-lingual learning without additional tags to distinguish the languages. In addition, a strategy similar to ELMo (Peters et al., 2018) is adopted, where a weighted sum of the outputs of all layers is computed as follows and fed to a task-specific classifier:

$$e_j^{task} = \textstyle\sum_i mBERT_{ij}.$$

Here, e^{task} denotes the contextual output embeddings for tasks such as the dependency parse. In addition, $mBERT_{ij}$ denotes the $mBERT$ representation for layer i at token position j.

In the task involving dependency structures, mBERT's subword tokenization process inputs words into multiple subword units. However, only the embeddings e_j^{task} of the first subword unit are used, serving as input to the graph-based bi-affine attention classifier (Dozat and Manning, 2017). The resulting outputs are combined using bi-affine attention to produce a probability distribution of the arc-head for each word. Finally, the dependency tree is decoded using the Chu–Liu/Edmonds algorithm (Chu, 1965; Edmonds et al., 1967).

2.2. Unsupervised Dependency Learning

Adhering to the properties of dependency syntax (Robinson, 1970), a general unsupervised algorithm for projective N-gram dependency learning (Unsupervised-Dep) was described in Ding and Yamamoto (2013, 2014). This method constructs the best dependency tree with a dynamic programming method using a CYK style chart and is based on the complete-link and complete-sequence nonconstituent concepts. However, considering the time complexity of this approach for arbitrary N-gram dependency learning, which may not be ideal for practical applications, we chose to focus in this study on the case of the bi-gram.

When considering the bi-gram, the directionality of a pair of words is set by the dependency relation, with $(w_i{\rightarrow}w_j)$ indicating a rightward relation and $(w_i{\leftarrow}w_j)$ indicating a leftward one. The bi-gram unsupervised learning update probabilities $P(w_i{\rightarrow}w_j)$ and $P(w_i{\leftarrow}w_j)$ are calculated using the Inside–Outside algorithm (Lari and Young, 1990). Finally, the Viterbi algorithm (Forney, 1973) is employed to determine the tree construction in the calculated Inside portion with the maximum probability, thus generating the optimal structure.

3. Investigation

3.1. UDify with Data Augmentation

In the work of Mao et al. (2023), a data augmentation based on Unsupervised-Dep is provided. Due to Unsupervised-Dep has a high time complexity of $O(n^3)$, making the common practice in the original methods, which start training from a random probability, somewhat inefficient. To circumvent this, the parsing results from UDify were utilized to initialize the probabilities. Despite the potential decrease in UDify's accuracy on low-resource languages during its training, the final results consistently outperform those from other parsing models (Qi et al., 2018; Tran and Bisazza, 2019), providing a solid foundation for the proposed initialization approach.

The process starts with the raw corpus, $Data$, input into the trained UDify by the original UD treebank, to generate the dependency arc-heads, represented as DEP_{arc}, and POS, lemmas, etc., denoted as $Others$. Statistical computations on DEP_{arc} generate initial probabilities $P(w_i{\rightarrow}w_j)$ and $P(w_i{\leftarrow}w_j)$, serving as input for Unsupervised-Dep alongside $Data$.

Following several iterations of training through Unsupervised-Dep, the re-estimated $P(w_i{\rightarrow}w_j)^{'}$ and $P(w_i{\leftarrow}w_j)^{'}$ emerge. They become the parameters for the Viterbi algorithm to determine the optimal dependency arc-head as given by

$$DEP_{arc}^{'} = Viterbi(x, P(w_i{\rightarrow}w_j)^{'}, P(w_i{\leftarrow}w_j)^{'}) \ ,$$

where $DEP_{arc}^{'}$ is the tree with the highest probability for a sentence x from $Data$.

Finally, $DEP_{arc}^{'}$ is merged with $Others$, ultimately generating artificial data. The artificial data are then combined with the existing UD treebanks for the subsequent UDify training.

3.2. On Few- and Zero-Shot Languages

During the training of UDify, the dependency structures for zero-shot languages are learned through

[1] github.com/google-research/bert/multilingual.md

transfer learning. Compared to high-resource languages, an early saturation in the accuracy of dependency parsing is observed across all zero-shot languages during the learning process. The peak performance is typically reached around the 12th training epoch, as illustrated in Figure 2. Mao et al. (2023) applied data augmentation to individual zero-shot language, effectively addressing this issue.

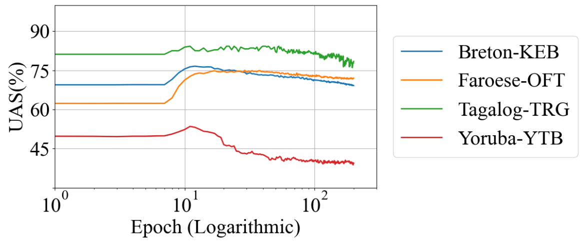


Figure 2: Change in the UAS(%) of low-resource languages during UDify(our) training.

However, when applying Unsupervised-Dep data augmentation to multiple zero-shot languages, the effectiveness of this approach has not been explored due to the impact of the amount of data generated on parser performance. Especially considering that this approach may generate large amounts of artificial data, its practical application in this context needs to be evaluated.

Moreover, the training of multilingual parser reveals that few-shot languages are similarly affected by the volume of training data. This highlights the critical need for effective data augmentation methods to improve the parsing performance of models like UDify. We aim to employ Unsupervised-Dep for multiple languages to explore its potential in mitigating early saturation in zero-shot languages and improving parsing accuracy in few-shot languages within a multilingual context.

4. Experiments

4.1. Dataset

The raw data of seven few-shot and four zero-shot languages that are most often tokenized using spaces were collected from El-Kishky et al. (2020); Fan et al. (2021); Schwenk et al. (2021) to create our selected low-resource language set for the implementation of Unsupervised-Dep. The data in the experiment are summarized in Table 1 and referred to as OPUS-mult in subsequent sections.

For comparison with the UDify and to illustrate our motivation, our parser experiments employed the UD Treebank v2.3 used by UDify. During training, following McDonald et al. (2011), we merged training sets, randomized the sentence order each epoch, and fed the network diverse batches of original and artificial data from multiple languages.

language(code)	#sent.(len.)	#train	#test
Armenian(hy)	2.4(8.2)	560	470
Belarusian(be)	2.0(9.0)	260	68
Hungarian(hu)	134.1(5.3)	910	449
Kazakh(kk)	1.7(8.2)	31	1,047
Lithuanian(lt)	236.7(5.6)	153	55
Marathi(mr)	1.5(10.0)	373	47
Tamil(ta)	13.7(7.7)	400	120
Breton(br)	18.2(9.5)	0	888
Faroese(fo)	1.3(8.1)	0	1,208
Tagalog(tl)	150.0(16.2)	0	55
Yoruba(yo)	9.7(8.1)	0	100

Table 1: Raw data collected from various corpora. Above: few-shot languages; below: zero-shot languages. #sent.(len.) denotes the raw sentences in unsupervised learning (in thousands), with the numbers in parentheses indicating the average length. #train and #test are the sentence counts in UD v2.3 treebank's training and test sets, respectively.

4.2. Setup

To minimize the impact of experimental environment variations on the result of Popel and Bojar (2018) in the comparisons, we followed the parameter settings from Kondratyuk and Straka (2019) and re-implemented the model as UDify(our). Additionally, to expedite the training process, we employed Horovod (Sergeev and Balso, 2018) to implement parallel computation.

At the beginning of training on Unsupervised-Dep, we used the UDify(our) model to parse each language present in the OPUS-mult dataset. The statistical results derived from the parsing outcomes of each language were adopted as its initial probabilities, which were continuously re-estimated throughout the unsupervised learning process. After the 10th training iteration, we employed the newly estimated probabilities to parse the sentences from OPUS-mult.

To assess the impact of augmenting training data for multiple low-resource languages on parsing accuracy, we designed and conducted several experiments. In the Unsupervised-Dep data augmentation experiments, we randomly selected 300 sentences for each language from OPUS-mult, processed them using Unsupervised-Dep, and integrated them into the UD treebanks to form the training dataset. The model trained from this dataset is referred to as Unsup. Inspired by the work of Rybak and Wróblewska (2018), we conducted a comparative experiment using a data augmentation method dubbed Self. In this approach, we used the same raw sentences train Unsup model and directly applied the parsing results obtained from the UDify(org) model. These results were merged with the original training set to train the Self model.

	hy	be	hu	kk	lt	mr	ta	br	fo	tl	yo	Few	Zero
UDify(org)	85.6	91.8	89.7	74.8	79.1	**79.4**	79.3	63.5	67.2	64.0	37.6	-	-
UDify(our)	86.1	92.1	89.8	76.0	79.4	74.3	80.8	69.2	72.0	78.4	39.4	84.0	67.1
Self	85.9	**92.5**	89.6	**76.2**	79.2	74.8	**81.2**	69.8	**72.5**	85.3	38.8	84.0	67.6
Unsup	**86.3**	92.4	**90.0**	**76.2**	**79.5**	74.0	80.5	**72.7**	71.9	**88.0**	**39.6**	**84.2**	**68.7**

Table 2: UAS(%) for few- and zero-shot languages obtained using different methods. The last two columns display the combined test set results for few- (Few) and for zero-shot (Zero) languages. We denote the treebank names using language codes; both the low-resource languages have only one treebank in UD v2.3. The UDify(org) result was reported in Kondratyuk and Straka (2019).

	Zero-shot		Other	
	UAS	Rest	UAS	Rest
UDify(our)	67.1	55.6	77.5	82.5
Self	67.6	56.3	77.5	82.4
Unsup	**68.7**	**59.0**	77.5	82.5

Table 3: UD scores on selected zero-shot and other languages obtained by different methods. Rest(%) refers to the average score of UPOS, UFeats, Lemma, and LAS in the UD scores.

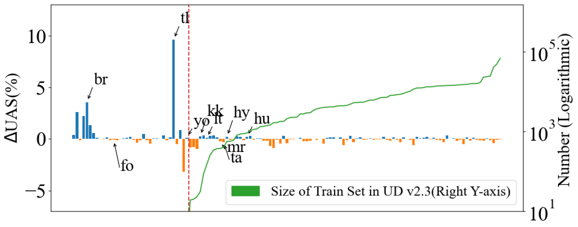


Figure 3: Difference in UAS(%) on all test treebanks: blue indicates Unsup > UDify(our), orange indicates Unsup < UDify(our). The left side of the red dotted line shows zero-shot languages.

4.3. Result and Discussion

A comparison with the experimental findings from Kondratyuk and Straka (2019) confirms the successful re-implementation of UDify(our), as illustrated in Table 2, and reveals that our replicated model surpasses those in related work (Choudhary, 2021; Üstün et al., 2022; Mao et al., 2023). Although no method produced a noticeable improvement for the few-shot languages, the results in this table indicate a significant improvement in UDify's ability to parse the dependency arc-head accuracy for zero-shot languages at the end of the training with the Unsupervised-Dep data augmentation method. This is reflected in the results for the combined test set, where the UAS increased to 68.7%. Taking Breton from the zero-shot languages as an example, we illustrate the changes in UAS during the training process under different methods in Figure 1. The figure reveals that the inclusion of data generated through Unsupervised-Dep significantly mitigates the reduction in UAS accuracy for zero-shot languages over the course of the training, thereby improving the result.

The UAS of almost every zero-shot language improved when artificial data via Unsupervised-Dep were included. To our knowledge, this is the state-of-the-art result for Tagalog. The `Tagalog-TRG` treebank is quite small, encompassing only 55 sentences with an average sentence length of 4.2 words in UD v2.3. In contrast, we have gathered 150k Tagalog raw sentences with an average length of 16.2 words. We believe that the quality and quantity of raw sentences used for training Unsupervised-Dep have a crucial impact on the performance of the multilingual parser.

To further enhance UDify's dependency parsing accuracy in low-resource languages, we attempted to increase the number of sentences generated by Unsupervised-Dep data augmentatio to 500, which we refer to as Unsup$^+$. In the result of Unsup$^+$, the UAS of the selected zero-shot languages in the test set saw further improvement, reaching 69.3%. We depict the changes in UAS for Breton during the Unsup$^+$ training process in Figure 1.

Given UDify's standing as a multilingual and multi-task parser, assessing the impact of our proposed methods on other languages and tasks is essential. To further scrutinize the variations between the UAS results of UDify(org) and Unsup, we carried out tests on all treebanks. As shown in Figure 3, the results indicate that Unsup effectively enhanced the UAS of zero-shot languages when artificial data were created using Unsupervised-Dep, especially for Breton and Tagalog. Meanwhile, its impact on the parsing precision of dependency structures in other languages is negligible.

For a comprehensive comparison, the UD scores of the zero-shot and other languages have been compiled in Table 3. Given that UDify must balance the loss produced by multiple decoders during training and the work of Rybak and Wróblewska (2018), these variations in evaluation metrics are considered reasonable. Broadly, our method has not had a negative impact on other languages and tasks, maintaining their performance levels.

Considering all results, we argue that creating training data for multiple low-resource languages using Unsupervised-Dep is both essential and effective in multilingual modeling contexts.

5. Conclusion and Future Work

This study highlights the issue of early saturation in parsing accuracy for UDify across multiple low-resource languages. To address this challenge, we implemented data augmentation for several low-resource languages through unsupervised learning. The experimental results demonstrated the effectiveness of data augmentation method in enhancing the parsing performance of multilingual parsers for low-resource languages.

Despite the limitations posed by training speed and the quality and quantity of raw data on our experiments, two possibilities remain: (1) Generating more data for zero-shot languages could lead to positive improvements. (2) The quality and quantity of raw data play a crucial role in the effectiveness of unsupervised data augmentation methods, thereby affecting the performance of multilingual parsers.

In future work, our research aims to explore additional influencing factors and considerations to further enhance multilingual parsing performance in low-resource language scenarios. Moreover, we plan to conduct research and exploration on low-resource languages using the latest UD treebanks.

6. Bibliographical References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 2021. Multilingual AMR parsing with noisy knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2778–2789, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chinmay Choudhary. 2021. Improving the performance of UDify with linguistic typology knowledge. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 38–60, Online. Association for Computational Linguistics.

Yoeng-Jin Chu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenchen Ding and Mikio Yamamoto. 2013. An unsupervised parameter estimation algorithm for a generative dependency n-gram language model. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 516–524, Nagoya, Japan. Asian Federation of Natural Language Processing.

Chenchen Ding and Mikio Yamamoto. 2014. A generative dependency n-gram language model: Unsupervised parameter estimation and application. *Information and Media Technologies*, 9(4):857–885.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jack Edmonds et al. 1967. Optimum branchings. *Journal of Research of the national Bureau of Standards B*, 71(4):233–240.

Thomas Effland and Michael Collins. 2023. Improving low-resource cross-lingual parsing with expected statistic regularization. *Transactions of the Association for Computational Linguistics*, 11:122–138.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

G David Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Karim Lari and Steve J Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer speech & language*, 4(1):35–56.

Jiannan Mao, Chenchen Ding, Hour Kaing, Hideki Tanaka, Masao Utiyama, and Tadahiro Matsumoto. 2023. Improving zero-shot dependency parsing by unsupervised learning. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 217–226, Hong Kong, China. Association for Computational Linguistics.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Martin Popel and Ondrej Bojar. 2018. Training tips for the transformer model. *Prague Bull. Math. Linguistics*, 110:43–70.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Jane J. Robinson. 1970. Dependency structures and transformational rules. *Language*, 46(2):259–285.

Piotr Rybak and Alina Wróblewska. 2018. Semi-supervised neural system for tagging, parsing and lematization. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 45–54, Brussels, Belgium. Association for Computational Linguistics.

Manon Scholivet, Franck Dary, Alexis Nasr, Benoit Favre, and Carlos Ramisch. 2019. Typological features for multilingual delexicalised dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3919–3930, Minneapolis, Minnesota. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Alexander Sergeev and Mike Del Balso. 2018. Horovod: Fast and easy distributed deep learning in TensorFlow. *arXiv preprint arXiv:1802.05799*.

Ke Tran and Arianna Bisazza. 2019. Zero-shot dependency parsing with pre-trained multilingual sentence representations. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 281–288, Hong Kong, China. Association for Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2022. UDapter: Typology-based language adapters for multilingual dependency parsing and sequence labeling. *Computational Linguistics*, 48(3):555–592.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Part-of-Speech Tagging for Northern Kurdish

Peshmerge Morad[1] **Sina Ahmadi**[2] **Lorenzo Gatti**[3]
[1,3]University of Twente [2]University of Zurich
[1]p.morad@hotmail.com [2]sina.ahmadi@uzh.ch [3]l.gatti@utwente.nl

Abstract

In the growing domain of natural language processing, low-resourced languages like Northern Kurdish remain largely unexplored due to the lack of resources needed to be part of this growth. In particular, the tasks of part-of-speech tagging and tokenization for Northern Kurdish are still insufficiently addressed. In this study, we aim to bridge this gap by evaluating a range of statistical, neural, and fine-tuned-based models specifically tailored for Northern Kurdish. Leveraging limited but valuable datasets, including the Universal Dependency Kurmanji treebank and a novel manually annotated and tokenized gold-standard dataset consisting of 136 sentences ($2,937$ tokens). We evaluate several POS tagging models and report that the fine-tuned transformer-based model outperforms others, achieving an accuracy of 0.87 and a macro-averaged F1 score of 0.77. Data and models are publicly available under an open license at https://github.com/peshmerge/northern-kurdish-pos-tagging

Keywords: Part-of-Speech tagging, morphosyntactic analysis, Northern Kurdish, low-resource NLP

1. Introduction

Automatic part-of-speech (POS) tagging or grammatical tagging is the process of assigning POS tags to each word/token in a given text. POS tagging is essentially a disambiguation task because words naturally are ambiguous and can have more than one correct tag depending on the context and their position in the sentence. POS tagging serves many purposes in natural language processing (NLP) applications, and it is traditionally considered a building block for other tasks such as named entity recognition (Ma and Liu, 2021), information extraction (Luan et al., 2017), spelling correction (Nagata et al., 2018), text classification (Pranckevičius and Marcinkevičius, 2016), natural language generation (Li et al., 2019), and machine translation (Hlaing et al., 2022).

Just as part-of-speech tagging serves as a precursor for tasks like syntactic parsing, tokenization is a crucial task in NLP and a prerequisite for POS tagging. Tokenization is segmenting the input text into smaller, distinct units termed *tokens*. These tokens can encompass compound words, single words, sub-words, symbols, or other significant elements. At its most fundamental level, tokenization separates tokens using whitespace as a delimiter (Mitkov, 2022, p. 549).

Unlike high-resourced languages (HRLs) like English and French, for which POS tagging and tokenization have been extensively addressed, low-resourced languages (LRLs) like Kurdish lack sufficient tools and resources (Ahmadi, 2020a). Although Northern Kurdish is included in Universal Dependencies (UD) (Nivre et al., 2020) (using the 'Kurmanji' label since version 2.1) based on Gökırmak and Tyers (2017)'s treebank, hence serving as a benchmark, achieving high-accuracy POS tagging for LRLs may require a greater emphasis on linguistic insights as observed in other languages (Manning, 2011). Our literature review indicates that there is room for effective and open-source contributions to Kurdish POS tagging.

In this paper, we report on the progress we have made in addressing the task of POS tagging for Northern Kurdish. More specifically, we revisit the UD Kurmanji treebank (Gökırmak and Tyers, 2017) by reannotating tokens that belong to specific word classes and introducing a different annotation scheme with more fine-grained linguistic features of Northern Kurdish. Secondly, we create a manually tokenized and annotated gold-standard dataset for Northern Kurdish with a total of 136 sentences and $2,937$ tokens. To that end, we deploy an annotation scheme different from that of UD Kurmanji that aims for a more fine-grained representation of linguistic features of Northern Kurdish, notably noun phrases containing *Izafe* (also spelled *Ezafe*) acting as a relativizer and linker. Thirdly, we evaluate the effect of different POS techniques along with the annotation schemes. Finally, we implement different POS tagging models and introduce a state-of-the-art transformer-based POS tagger for Northern Kurdish.

The rest of the paper is organized as follows. In section 2, we provide an overview of the Kurdish language and its dialects, focusing on Northern Kurdish. Section 3 presents a comprehensive review of related work and state-of-the-art studies on POS tagging for LRLs in general, with a specific focus on Northern Kurdish. We then detail the annotation schemes for the training and testing datasets in section 4. In section 5, we discuss the process of collecting and annotating testing data, as well as augmenting the training data. Additionally, we provide a detailed explanation of the tokenization and POS tagging methods. Subsequently, section 6 presents our evaluation results, accompanied by an in-depth analysis. Finally, our conclusions are presented in section 7.

2. Kurdish Language

The Kurdish language belongs to the Northwestern Iranic branch within the Indo-European languages family, spoken by more than 30 million people. The Kurdish language (ISO 639-3 code kur) is divided into many dialects (with corresponding ISO 639-3 languages codes): Northern Kurdish or Kurmanji (kmr), Central Kurdish or Sorani (ckb), Southern Kurdish (sdh), and Laki (ldk) and is closely related to Zaza-Gorani languages (Ahmadi et al., 2019). Northern Kurdish is widely spoken in Syria and Turkey but also in the Kurdistan Region of Iraq, Iran, Armenia and among the Kurdish diaspora. It is written using Kurdified Latin-based and Arabic-based scripts. The Latin-based script is widely known as the Hawar alphabet introduced by Jeladet Ali Bedirkhan in 1932.

Northern Kurdish has a subject–object–verb word order and specifies grammatical gender (feminine and masculine). The noun in its absolute state and without any suffixes represents the generic and definite senses of the noun, and it marks four cases, namely nominative, oblique, *Izafe*, and vocative. In addition, it has a split-ergative alignment in the past tense with transitive verbs. Furthermore, the passive voice (conjugated in all persons, moods, and tenses) is constructed using the verb *hatin* 'to come' and *dan* 'to give' plus the infinitive.

Both the oblique and the *Izafe* case (construct case) are essential in Northern Kurdish for indicating the roles of the nouns and the pronouns in a sentence. Nouns, proper nouns, personal pronouns, and demonstrative adjectives, in both cases, undergo a form change as in "*komputera min*" (my computer) where 'a' is an *Izafe* linking '*komputera*' (computer) to '*min*' (my). They are either completely altered, or the case markers are added to the end of the noun and proper nouns. Those markers, shown in Table 1, are unstressed markers that reveal the gender and number of nouns. In this study, our introduced annotation scheme, discussed in Section 5.1, particularly revolves around addressing and segmenting the oblique and *Izafe* case markers in our datasets.

Nonetheless, *Izafe* case markers differ from oblique case markers in the fact that they can also appear as separate particles serving the same purpose within definite nouns; this phenomenon is referred to as *construct extender* (Thackston, 2006) because it allows extending the *Izafe* case by adding adjectives or nouns to the first *Izafe* case.

	OBLIQUE		IZAFE	
	Definite	Indefinite	Definite	Indefinite
SG. F.	*-ê*	*-ekê*	*-a*	*-eke/-eka*
SG. M.	*-î*	*-ekî*	*-ê*	*-ekî*
PL	*-an*	*-inan*	*-ên*	*-ine*

Table 1: Case markers based on the number, gender, and definiteness of the noun in Northern Kurdish. If the noun ends in a vowel, the case markers will be preceded by a *-y*.

3. Related Work

The task of POS tagging has been addressed using various methods. Rule-based techniques (Brill, 1992; Karlsson, 1990) were the first methods applied. Decision Trees have also been employed for the task (Schmid, 1994). Furthermore, hidden Markov models (HMMs) and conditional random fields (CRFs) have been widely used and proved to be effective for this task (Schmid and Laws, 2008; Pradhan and Yajnik, 2023; Yousif, 2019; Stratos et al., 2016; Silfverberg et al., 2014).

Additionally, deep learning based approaches like recurrent neural networks and (Bi)LSTMs have shown to be powerful in capturing temporal dependencies when performing POS tagging (Wang et al., 2015; Qi et al., 2020; Horsmann and Zesch, 2017). Those are often combined with other techniques such as convolutional neural networks, HMMs, and CRFs (Shao et al., 2017; Plank et al., 2016; Ma and Hovy, 2016; Maimaiti et al., 2017).

In recent years, the rise of transformer-based architectures introduced by Vaswani et al. (2017) has led to the development of large language models (LLMs) such as GPT2 (Radford et al., 2019), BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019). These models have greatly influenced NLP in various fields. However, despite being trained on multiple languages, they don't always perform better than single-language models, especially in less-resourced languages, for tasks like POS tagging (Conneau et al., 2020). Nonetheless, they can adapt and improve their performance when fine-tuned (Maimaiti et al., 2021).

For Kurdish, Walther et al. (2010) presents the first dedicated work on POS tagging for Northern Kurdish, where a morphological lexicon (KurLex) and a POS tagger were created. The authors report an 85.7% precision, however on a small annotated corpus of 13 sentences. Although Gökırmak and Tyers (2017)'s treebank for Northern Kurdish is available on UD and has been used in various consecutive studies in multilingual training setups as in Qi et al., 2020 (BiLSTM) and Nguyen et al., 2021 (transformer-based fine-tuning) inter alia, there is still no tool or fine-grained dataset indicating the existing gap in the literature (Ahmadi, 2020a).

4. Annotation Schemes

4.1. UD Kurmanji Scheme

The UD Kurmanji treebank (Gökırmak and Tyers, 2017) is a treebank for Northern Kurdish that contains morpho-syntactic information such as POS tags and some morphological features. The data in the treebank is drawn from fiction and encyclopedic data in roughly equal measure. It consists of the Kurdish translation of *The Adventure of the Speckled Band* story and sentences from the Northern Kurdish Wikipedia. UD Kurmanji contains $10,189$ tokens and has been annotated following the UD annotation scheme (Nivre et al., 2020), meaning it does not allow multi-word expressions, and it instructs to undone contractions. In addition, the case markers, shown in Table 1, within nouns are not segmented. Moreover, the construct extenders in the treebank are tagged as ADP. For example, the noun phrase '*Beşa Felsefeyê*' (department of philosophy) is tagged as NOUN and NOUN, respectively, while having *Izafe* and oblique case markers in both nouns.

4.2. Our Scheme

We propose a different, fine-grained annotation scheme taking into account all case and indefinite noun markers. In addition, we address multi-word prepositions such as 'lê' (from, analogous to *au/aux* in French), adverbs, and compound verbs and tag them as single tokens. It is worth mentioning that the UD annotation scheme (Nivre et al., 2020) serves as a basis for our scheme.

Case Markers and Determiners One of the main differences between our scheme and the UD Kurmanji scheme is how we segment the nouns and their attached indefinite, oblique, and *Izafe* case markers. We use the POS tags from the UD tagset (Petrov et al., 2012). While we use DET for indefinite and oblique case makers, we introduce a new POS tag named IZAFE for the *Izafe* case markers. For example, the noun phrase *Beşa Felsefeyê* (department of philosophy) is split into four tokens *Beş*, *a*, *Felsefe yê* and respectively tagged as NOUN, IZAFE, NOUN and DET.

Multi-word Expressions In UD Kurmanji, the tag X is assigned to nouns that are part of the compound verbs; in our case, we tag those nouns either as a NOUN or all together with the verbs they belong to as a multiword expression VERB. For instance, in UD Kurmanji, the compound verb '*pêşkêş dikin*' (presenting) is split into two tokens: *pêşkêş* and *dikin* and tagged X and VERB, respectively. Within our annotation scheme, we tag it as VERB.

Regarding compound prepositions, we annotate the compound preposition '*li ser*' (on/upon) as ADP, while in UD Kurmanji, it is separated into two tokens '*li*' (in/at) and '*ser*' (onto) where both are tagged as ADP. In addition, compound adverbs such as '*bi tenê*' (only) are also separated into two tokens '*bi*' (with) and '*tenê*' (alone), both are annotated as ADP. However, we treat it as a multi-word token, and we annotate it as ADV.

Moreover, the verb to be in Northern Kurdish '*bûn*' (to be) is always annotated as AUX in UD Kurmanji treebank, while we tag it as a VERB unless it appears as a light verb. In addition, the particles *-ê* and *dê* are used for forming the future tense in Northern Kurdish and are tagged as AUX in UD Kurmanji. However, we tag those particles as PART because they are not auxiliary verbs.

Furthermore, the tokens '*jî*' (also/too) and '*her*' (every) are annotated as PART and either as DET or ADV in the UD Kurmanji, respectively. We annotate the former as ADV and the latter as PRON.

5. Methodology

5.1. Data Collection and Annotation

We collect 136 ($2,937$ tokens) sentences written in Northern Kurdish from multiple news websites. The first 100 sentences are taken from the unannotated Pewan corpus (Esmaili et al., 2013). The remaining 36 sentences are taken from three Kurdish news websites, mainly Kurdistan24[1], Xwebûn[2], and Hawar News[3]. We annotated those sentences according to our annotation scheme introduced in section 4.2. We call this collection the ”gold-standard dataset”, and we use it as a test set to evaluate our POS tagging models. Figure 1 demonstrate the statistics of this dataset.

Similar to the UD Kurmanji treebank, for each given sentence in our gold-standard dataset, we provide: 1) the raw (untokenized) sentence where tokens are delimited by whitespaces and the case markers are not split-off, and 2) a list of tokens with corresponding POS tags where the case markers are segmented and annotated.

The availability of the untokenized sentence, along with the list of the tokens, enables us to evaluate various tokenization methods. The untokenized sentence can be fed to any tokenizer, and its output can be compared against the list of tokens we already have, which we consider as gold tokens.

[1] https://www.kurdistan24.net/kmr

[2] https://xwebun1.org

[3] https://hawarnews.com/kr

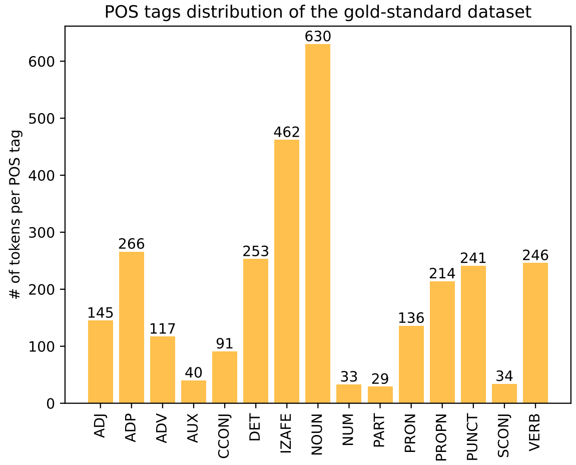


Figure 1: Number of tokens per POS tags in our gold-standard dataset

5.2. Data Augmentation

We augment the UD Kurmanji treebank by splitting the case and indefinite markers from the tokens they are attached to. Thus introducing new tokens. For example, we split '*hevalekî*' (a male friend) into three separate tokens, each with its corresponding POS tag: *heval* as `NOUN`, *ek* (indefinite noun marker) as `DET`, and finally *î* as `IZAFE`. In addition, we re-tag independent *Izafe* markers (construct extender) as `IZAFE` instead of `ADP`. Finally, we reverse the splitting of the contracted prepositions (*jê, lê, pê, tê*) in the treebank.

Our approach for augmenting the UD Kurmanji treebank bears a close resemblance to the research described by (Seddah et al., 2023). The authors made significant steps in addressing tokenization issues to ensure consistency in the NArabizi treebank annotations (Farah et al., 2020), the user-generated content variety of Arabic Algerian, which is known for its frequent usage of code-switching. For instance, they carefully segmented specific classes of words, such as determiners in noun phrases.

As a result of this augmentation step, the number of tokens increased in the treebank ($12,233$ tokens). We refer to this augmented version as UD Kurmanji augmented, while we refer to the version with its initial annotation scheme as UD Kurmanji original.

5.3. Tokenization

In addition to the KLPT tokenizer, Ahmadi, 2020b provided multiple neural tokenization models trained (unsupervised) on Northern Kurdish raw corpora. We use three of those models: Unigram (Kudo, 2018), Byte-Pair Encoding (BPE) (Sennrich et al., 2016), and wordPiece (Schuster and Nakajima, 2012) tokenizers.

Moreover, we use the NLTK tokenizer and a manual tokenization method. The manual tokenization, as the name suggests, is the process of manually tokenizing any given text. This method is mostly performed in pairs with the task of manually annotating tokens with the corresponding POS tags. Despite being very time-consuming, it is considered to have the best outcome because it is done by humans with good linguistic knowledge of the language. Therefore, the manually tokenized text can be considered the ground truth that can be used for evaluating other automatic tokenization methods.

5.4. POS Tagging

The task of POS tagging can be seen as a multi-class classification task where a model is trained on annotated data to enable it to classify each token in any given sequence of tokens. There are multiple approaches to tackle the task of POS tagging. Generally, those approaches can be grouped into four categories: rule-based, statistical, neural-based, and transformer-based fine-tuned (Jurafsky and Martin, 2009; Kanakaraddi and Nandyal, 2018).

Except for the work of Walther et al., 2010, there has been no dedicated work for the task of POS tagging for Northern Kurdish. Therefore, we propose seven supervised POS tagging models. The goal is to cover POS methods as much as possible to establish a baseline method and to examine the effectiveness of those methods. Those methods will be explained in the following subsections.

It is worth mentioning that we train all POS tagging models independently, once on the UD Kurmanji original and once on the UD Kurmanji augmented. We take this approach because we want to assess the impact of the annotation scheme on the models' performance. Hence, the labels (augmented) and (original) within the models' names indicate the dataset used for training the model, either UD Kurmanji augmented or UD Kurmanji original.

5.4.1. Statistical-based Models

Our first model is a Unigram model from the NLTK Python package (Bird et al., 2009). This model assigns tags based on word frequency observed during training. It uses conditional frequency distributions to calculate the most likely tag for each given token. The model may encounter unfamiliar words in linguistically resource-limited settings like ours (*out-of-vocabulary*). Therefore, we specify the default POS tag as `NOUN` when it fails to determine a POS tag for a token. This is a common practice when establishing a baseline, and it is motivated by Bird et al. (2009).

In addition, we create HMM (Huang et al., 2001) and CRF (based on CRFsuite library (Okazaki, 2007)) models using the implementation available in the NLTK Python package.
Finally, we create an ExtraTrees POS model using the implementation from Scikit-learn (Pedregosa et al., 2011).

5.4.2. Neural-based Models

Our first neural-based model is the Averaged Perceptron POS tagging model, similar to the Extra Trees model, which has the notion of feature engineering. However, here we do not define our own set of features, we use the standard features set defined by the NLTK Python package since we use their implementation[4].

In addition, we use the Flair Python package (Akbik et al., 2019) to create a BiLSTM model using a configurable BiLSTM architecture as originally proposed by Huang et al. (2015). For this model, we use pre-trained sub-word fastText embeddings (Grave et al., 2018) specifically pre-trained on Northern Kurdish data. FastText enables us to generate embeddings from character-level n-grams, thereby being better at capturing morphological nuances.

5.4.3. Transformer-based Fine-tuned Models

In contrast to the previous models, where each model was trained from scratch for our task, we fine-tune the pretrained multilingual XLM-RoBERTa model (Conneau et al., 2020) on the UD Kurmanji original and UD Kurmanji augmented. We utilize the 'base' version of XLM-RoBERTa because of its lower computational requirements, making it easier to fine-tune. The fine-tuning is performed using Trankit (Nguyen et al., 2021), which offers a relatively fast and straightforward approach for fine-tuning LLMs like XLM-RoBERTa, thanks to the utilization of Adapters (Pfeiffer et al., 2020). We refer to the fine-tuned POS model as **N**orthern **K**urdish **XLM**-**R**oBERTa **(NK-XLMR)**.

6. Experiments

6.1. Tokenization Performance

We distinguish between two types of tokenization evaluation: 1) intrinsic evaluation and 2) extrinsic evaluation.
Within the intrinsic evaluation, we want to evaluate the quality of the tokenization system in isolation from the later stages, POS tagging in our case. The intrinsic evaluation directly measures the tokenization system's capabilities by comparing it to similar systems. We follow the same approach of (Ahmadi, 2020b) by performing tokenization evaluation using the Bilingual Evaluation Understudy Score (BLEU).

Table 2 shows the BLEU scores of the tokenization methods we used in this study using the gold-standard dataset as testing data. We see that the BLEU scores for the KLPT tokenizer are the highest, outperforming other tokenizers by a great margin. In contrast to other tokenizers, the KLPT tokenizer is characterized by its extensive knowledge of Northern Kurdish, enabling it to correctly recognize case markers and handle multi-word expressions like compound verbs and compound prepositions.

Tokenizer	BLEU-1	BLEU-2	BLEU-3	BLEU-4
KLPT	0.73	0.65	0.59	0.53
unigram	0.54	0.44	0.36	0.29
NLTK	0.50	0.41	0.33	0.25
BPE	0.50	0.39	0.31	0.24
wordPiece	0.45	0.36	0.28	0.21

Table 2: BLEU scores for all tokenization methods on the gold-standard dataset.

Within the extrinsic evaluation, we evaluate the tokenization system by measuring its impact on our whole NLP pipeline. In our case, the tokenization system's quality greatly affects the POS tagger's performance. Therefore, the tokenization correctness can also be determined by examining the F1 and accuracy scores of the POS tagger presented in section 6.2.

6.2. POS Tagging Performance

We present the evaluation results (accuracy and macro-averaged F1 score) of all POS tagging models. In order to make the comparison clearer, we divide the results based on the used training data (UD Kurmanji original and augmented). While table 4 provides a detailed comparison of all models trained on the UD Kurmanji augmented, table 3 demonstrates the results of the same POS model but trained on UD Kurmanji original.

By comparing the results in both tables and regardless of the tokenization method, we observe a performance increase among the models. This increase is the highest within the manual tokenization method and the lowest within the wordPiece tokenization method. This confirms the importance and the impact of the data augmentation we did on the UD Kurmanji original treebank for the task of POS tagging. In addition, it stipulates the impact the performance of the tokenization method has on

[4] This implementation is based on Matthew Honnibal's implementation: https://explosion.ai/blog/part-of-speech-pos-tagger-in-python

Model / Tokenizer	manual		KLPT		NLTK		unigram		BPE		wordPiece	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Baseline (Unigram)	0.4	0.51	0.35	0.41	0.37	0.32	0.36	0.34	0.36	0.32	0.34	0.31
HMM	0.37	0.46	0.33	0.36	0.35	0.32	0.34	0.33	0.34	0.32	0.34	0.31
ExtraTrees	0.41	0.52	0.37	0.42	0.38	0.33	0.37	0.34	0.38	0.33	0.34	0.32
AveragedPerceptron	0.44	0.54	0.37	0.42	0.40	0.36	0.38	0.37	0.39	0.35	0.36	0.33
BiLSTM	0.42	0.51	0.40	0.41	0.45	0.35	0.43	0.36	0.44	0.34	**0.42**	0.33
CRF	0.46	0.54	0.41	0.44	0.42	0.36	0.40	0.37	0.40	0.35	0.35	0.33
NK-XLMR	**0.57**	**0.62**	**0.46**	**0.47**	**0.47**	**0.38**	**0.45**	**0.39**	**0.45**	**0.37**	0.40	**0.35**

Table 3: The macro-averaged F1 scores and accuracy (Acc) of the POS tagging models trained on the UD Kurmanji original and evaluated on our gold-standard dataset.

Model / Tokenizer	manual		KLPT		unigram		NLTK		BPE		wordPiece	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Baseline (Unigram)	0.59	0.73	0.47	0.52	0.41	0.37	0.4	0.32	0.4	0.33	0.37	0.33
HMM	0.62	0.77	0.48	0.53	0.4	0.37	0.41	0.33	0.4	0.34	0.38	0.33
ExtraTrees	0.61	0.79	0.49	0.56	0.43	0.4	0.41	0.36	0.41	0.36	0.37	0.34
AveragedPerceptron	0.68	0.83	**0.57**	0.57	0.47	0.41	0.49	0.37	0.45	0.37	0.40	0.35
BiLSTM	0.72	0.83	0.52	0.57	0.45	0.40	0.43	0.36	0.43	0.37	0.41	0.34
CRF	0.74	0.84	0.55	0.59	0.48	0.42	0.48	0.39	0.46	0.38	0.42	0.35
NK-XLMR	**0.77**	**0.87**	0.56	**0.59**	**0.49**	**0.43**	**0.51**	**0.39**	**0.47**	**0.39**	**0.44**	**0.36**

Table 4: The macro-averaged F1 scores and accuracy (Acc) of the POS tagging models, trained on the UD Kurmanji augmented and evaluated on our gold-standard dataset.

POS tagging for Northern Kurdish. While this performance increase is in part due to the different annotation scheme, which is explained in section 4.2, the introduction of this richer scheme improved the performance of the POS models on specific POS tags other than IZAFE and DET. A detailed analysis of this improvement is reported in section 6.3.

Further observation reveals that within the context of the training on UD Kurmanji augmented, both the BiLSTM and AveragedPerceptron models exhibit identical accuracy scores, although their macro-averaged F1 scores diverge slightly but remain comparable. Conversely, when utilizing the UD Kurmanji original, a similar trend of identical accuracy emerges between the AveragedPerceptron and the CRF models. Additionally, it is notable that the HMM model falls behind, even when compared to the baseline.

Moreover, the NK-XLMR model is our best model as it outperforms all other models. This was an expected performance, and it is in line with our finding in section 3 where we showed how LLMs achieve state-of-the-art results for multiple NLP tasks, including POS tagging.

However, comparing the scores of NK-XLMR and CRF models in Table 4, we observe very close performance between the two. The difference is very small, 0.03 for the macro-averaged F1 and the accuracy scores. This is a notable result, especially with regard to the computational resources required for fine-tuning XLM-RoBERTa and for training the CRF model from scratch for the task of POS tagging. Based on our experiments in this study, fine-tuning XLM-RoBERTa for POS tagging took notably longer than training the CRF for the same task.

6.3. Analysis

The presented results in the Tables 4 and 3 unambiguously demonstrate two trends in our results. First, training the POS models on the UD Kurmanji augmented undeniably results in higher accuracy and F1 scores when compared with the outcomes of POS models trained on the UD Kurmanji original. Second, the performance of POS models tends to decline as we transition away from the manual tokenization method. The further we move, the less knowledge of the linguistic characteristics of Northern Kurdish the tokenizers have. To further analyze this, we present two confusion matrices in Figures 3a and 3b demonstrating the performance of the NK-XLMR(augmented) and NK-XLMR(original).

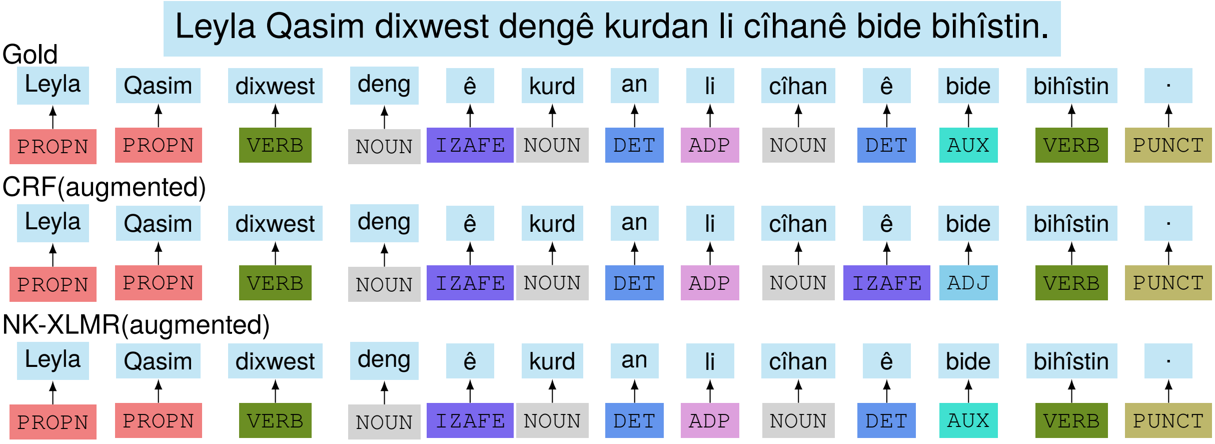
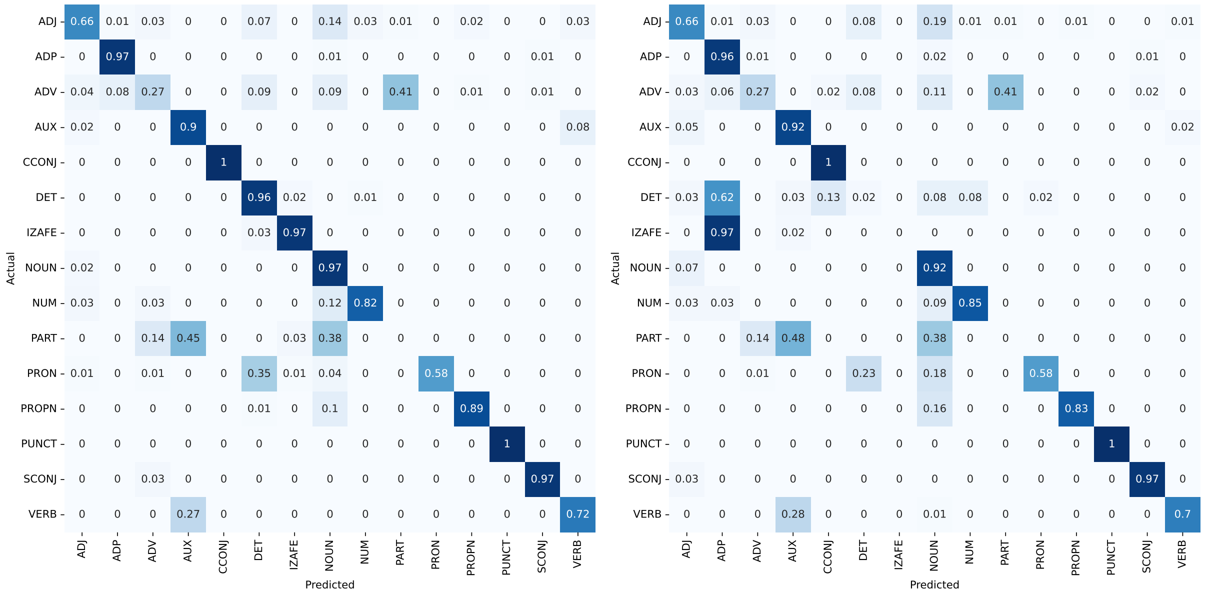


Figure 2: Outputs of the CRF and NK-XLMR compared to the gold annotations for a sentence from the gold-standard dataset (Translation: ‘Leyla Qasim wanted to make the Kurdish voice heard in the world.’)



(a) NK-XLMR (augmented)

(b) NK-XLMR (original)

Figure 3: Confusion matrices of NK-XLMR (augmented) and NK-XLMR (original) models. Although both models exhibit inadequacy in handling the PART and ADV tags, NOUN and PROPN benefit from data augmentation.

The UD Kurmanji augmented is characterized by the enhancements we have introduced and discussed in detail in section 5.2. The data augmentation affected tokens from the following POS tags: NOUN, PROPN, DET, and ADP, which are important elements in the *Izafe* and oblique cases in Northern Kurdish.

By comparing the confusion matrices, we observe that NOUN and PROPN benefit the most from the data augmentation, demonstrating 0.05 and 0.06 accuracy improvement, respectively, and the ADP and VERB to a lesser extent. In addition, we see that the tags DET and IZAFE enjoy huge improvement when trained on the UD Kurmanji augmented. However, we cannot consider it reliable since the IZAFE tag was not present in the UD Kurmanji original.

Nevertheless, it is evident that the NK-XLMR (original and augmented) exhibits a notable inadequacy in handling the PART and ADV tags. Examined outputs of NK-XLMR(augmented) and the error rates presented in section 6.3 and section 6.3 also verify this inadequacy. The tag PART has an error rate of 1.0, which means the model completely fails in recognizing tokens belonging to this tag correctly. We argue that this can be attributed to a misalignment in the annotation schemes between the UD Kurmanji and ours rather than a lim-

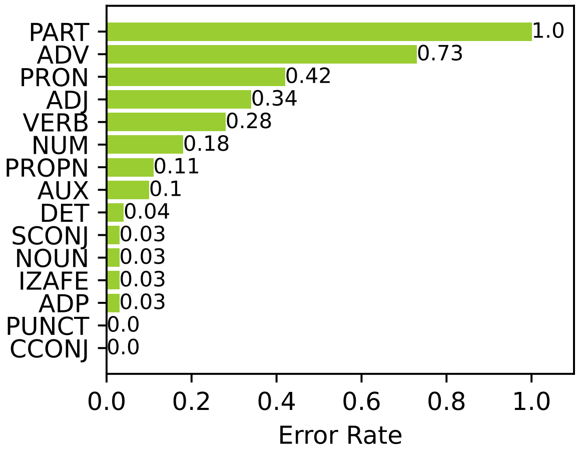


Figure 4: Error rates of NK-XLMR(augmented)

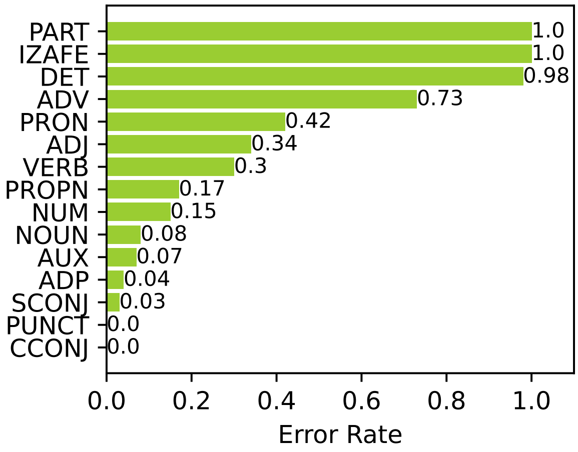


Figure 5: Error rates of NK-XLMR(original).

itation within the model itself.

Additionally, in section 6.3, we see that IZAFE and DET also have error rates of 1.0 and 0.98. This happens due to the fact that NK-XLMR(original) has no knowledge of the *Izafe* and oblique case markers and, therefore, fails to perform POS tagging correctly when evaluated on the gold-standard dataset where those markers are explicitly represented.

Regarding the second trend, the most straightforward reason for this is the fact that the tokenization methods are generating, in most cases, either fewer or more tokens than the ground truth. This can be attributed to the linguistic knowledge the tokenizer has about Northern Kurdish, such as the *Izafe*, oblique case markers, and multi-word expressions.

7. Conclusions and Discussion

The main objective of this study was to address the task of POS tagging for Northern Kurdish by utilizing the currently available resources. On the one hand, our multifaceted approach for this study enabled us to establish a baseline POS tagger for Northern Kurdish using the Unigram(augmented) model with an accuracy of 0.73 and a macro-averaged F1 score of 0.59 evaluated on the gold-standard dataset. On the other hand, the CRF(augmented) model achieves the second-best performance with 0.84 and 0.74 for accuracy and macro-averaged F1 score, making it the best-performing model among statistical POS tagging models. In addition, the CRF model stands out because of its quick training time.

The transformer-based NK-XLMR (augmented) outperforms all other models with an accuracy of 0.87 and a macro-averaged F1 score of 0.77., thus setting a new state-of-the-art performance for the task of POS tagging in Northern Kurdish. Our results are particularly robust compared to the work of Walther et al., 2010, where their POS tagger for Northern Kurdish was evaluated on only 13 sentences. This comparison underscores the reliability of our findings, considering the granularity of linguistic features in our gold-standard dataset and the larger number of test sentences (136 sentences) we used for evaluation.

Moreover, we further explored the impact of tokenization methods on POS tagging accuracy by comparing their outcomes against the gold standard tokens in our dataset. While encountering difficulties with certain linguistic nuances, the KLPT tokenizer demonstrated notable proficiency in capturing Northern Kurdish linguistic traits.

Finally, we successfully demonstrated the effect of the various linguistic features of Northern Kurdish, such as the *Izafe* and oblique case markers and contracted prepositions on the task by evaluating both variants of the models (original and augmented). Our POS tagging models trained on the UD Kurmanji augmented showed improvements on NOUN, PROPN, VERB, and ADP POS tags.

Limitations While this study has made several contributions to the field of Kurdish NLP, several limitations should be noted. Firstly, we did not target the task of syntactic parsing. Secondly, we did not explore the employment of LLMs or POS models from other closely related languages like Persian or dialects like Central Kurdish. Furthermore, we did not examine the impact of our POS tagging models and annotation schemes on other downstream tasks like named entity recognition, sentiment analysis , or parsing.

References

Sina Ahmadi. 2020a. KLPT–Kurdish language processing toolkit. In *Proceedings of second work-*

shop for NLP open source software (NLP-OSS), pages 72–84.

Sina Ahmadi. 2020b. A tokenization system for the Kurdish language. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 114–127.

Sina Ahmadi, Hossein Hassani, and John P McCrae. 2019. Towards electronic lexicography for the kurdish language. In *Proceedings of the sixth biennial conference on electronic lexicography (eLex)*. eLex 2019.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Eric Brill. 1992. A simple rule-based part of speech tagger. Technical report, Pennsylvania Univ Philadelphia Dept of Computer and Information Science.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Kyumars Sheykh Esmaili, Donya Eliassi, Shahin Salavati, Purya Aliabadi, Asrin Mohammadi, Somayeh Yosefi, and Shownem Hakimi. 2013. Building a test collection for Sorani Kurdish. In *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEEE.

Djamé Farah, Seddah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Ortiz Suarez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 1139–1150.

Édouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Memduh Gökırmak and Francis M. Tyers. 2017. A dependency treebank for Kurmanji Kurdish. In *Proceedings of the Fourth International Conference on Dependency Linguistics (DepLing, 2017)*, pages 64–73.

Zar Zar Hlaing, Ye Kyaw Thu, Thepchai Supnithi, and Ponrudee Netisopakul. 2022. Improving neural machine translation with POS-tag features for low-resource language pairs. *Heliyon*, 8(8).

Tobias Horsmann and Torsten Zesch. 2017. Do LSTMs really work so well for PoS tagging? – a replication study. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 727–736, Copenhagen, Denmark. Association for Computational Linguistics.

Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st edition. Prentice Hall PTR, USA.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Suvarna G Kanakaraddi and Suvarna S Nandyal. 2018. Survey on parts of speech tagger techniques. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pages 1–6. IEEE.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pretraining of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the*

56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 66–75.

Hailiang Li, YC Adele, Yang Liu, Du Tang, Zhibin Lei, and Wenye Li. 2019. An augmented transformer architecture for natural language generation tasks. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 1–7. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. *arXiv preprint arXiv:1708.06075*.

Liwen Ma and Weifeng Liu. 2021. An enhanced method for entity trigger named entity recognition based on pos tag embedding. In *2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pages 89–93.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Maihemuti Maimaiti, Aishan Wumaier, Kahaerjiang Abiderexiti, and Tuergen Yibulayin. 2017. Bidirectional long short-term memory network with a conditional random field layer for Uyghur part-of-speech tagging. *Information*, 8(4):157.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, Zegao Pan, and Maosong Sun. 2021. Improving data augmentation for low-resource NMT guided by POS-tagging and paraphrase embedding. *Transactions on Asian and Low-Resource Language Information Processing*, 20(6):1–21.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, pages 171–189. Springer.

Ruslan Mitkov. 2022. *The Oxford handbook of computational linguistics*. Oxford University Press.

Ryo Nagata, Tomoya Mizumoto, Yuta Kikuchi, Yoshifumi Kawasaki, and Kotaro Funakoshi. 2018. A POS tagging model adapted to learner English. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 39–48.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs). *URL http://www. chokkan. org/software/crfsuite*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2089–2096. European Language Resources Association (ELRA).

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418.

Ashish Pradhan and Archit Yajnik. 2023. Parts-of-speech tagging of Nepali texts with Bidirectional LSTM, Conditional Random Fields and HMM. pages 1–17. Springer.

Tomas Pranckevičius and Virginijus Marcinkevičius. 2016. Application of logistic regression with part-of-the-speech tagging for multi-class text classification. In *2016 IEEE 4th workshop on advances in information, electronic and electrical engineering (AIEEE)*, pages 1–5. IEEE.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Arij Seddah, Djamé Riabi, and Mahamdi Menel. 2023. Enriching the NArabizi treebank: A multifaceted approach to supporting an under-resourced language. In *The 17th Linguistic Annotation Workshop (LAW-XVII)@ ACL 2023*, page 266.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics (ACL).

Yan Shao, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2017. Character-based Joint Segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 173–183.

Miikka Silfverberg, Teemu Ruokolainen, Krister Lindén, and Mikko Kurimo. 2014. Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 259–264.

Karl Stratos, Michael Collins, and Daniel Hsu. 2016. Unsupervised part-of-speech tagging with anchor hidden markov models. *Transactions of the Association for Computational Linguistics*, 4:245–257.

Wheeler M Thackston. 2006. *Kurmanji Kurdish:A Reference Grammar with Selected Readings*. Harvard University.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Géraldine Walther, Benoît Sagot, and Karën Fort. 2010. Fast development of basic NLP tools: Towards a lexicon and a POS tagger for Kurmanji Kurdish. In *International conference on lexis and grammar*.

Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2015. Part-of-speech tagging with bidirectional long short-term memory recurrent neural network. *arXiv preprint arXiv:1510.06168*.

Jabar Yousif. 2019. Hidden Markov Model tagger for applications based Arabic text: A review. *Journal of Computation and Applied Sciences IJOCAAS*, 7(1).

Diachronic Analysis of Multi-word Expression Functional Categories in Scientific English

Diego Alves, Stefania Degaetano-Ortlieb, Elena Schmidt, Elke Teich
Saarland University
Saarbrücken, Germany
diego.alves@uni-saarland.de, s.degaetano@mx.uni-saarland.de
elsc00001@stud.uni-saarland.de , e.teich@mx.uni-saarland.de

Abstract

We present a diachronic analysis of multi-word expressions (MWEs) in English based on the Royal Society Corpus, a dataset containing 300+ years of the scientific publications of the Royal Society of London. Specifically, we investigate the functions of MWEs, such as stance markers ("it is interesting") or discourse organizers ("in this section"), and their development over time. Our approach is multi-disciplinary: to detect MWEs we use Universal Dependencies, to classify them functionally we use an approach from register theory, and to assess their role in diachronic development we use an information-theoretic measure, relative entropy.

Keywords: multi-word expressions, universal dependencies, relative entropy, discourse functions, diachronic analysis

1. Introduction

In this paper, we analyze multi-word expressions (MWEs) and the functions they fulfill in scientific writing, inspecting diachronic changes from the mid 17th century to today. From a communicative perspective, MWEs contribute to language efficiency as they constitute highly predictable linguistic material with a clear processing advantage for language users. Their use in scientific writing is particularly interesting due to the high informational load encountered within the scientific domain, where MWEs can act as devices to smooth the informational load in the signal (Conklin and Schmitt, 2012).

There has been a long-standing tradition to identify and analyze MWEs in scientific text and academic writing more widely, most prominently in research on English for Academic Purposes (EAP) (cf. Oakey (2020)). We combine this approach considering the Academic Formula List (AFL) with a UD-based approach, were we use the dependency relation label `fixed` to identify further MWEs not included in the AFL list. As it has been shown that scientific writing becomes increasingly conventionalized over time (see e.g.Degaetano-Ortlieb and Teich (2019)), the `fixed` MWEs are particularly important for a diachronic analysis aimed at investigating communicative efficiency. In this study, we focus on the most frequent grammaticalized fixed expressions identified from the RSC combined with a set of formulaic expressions commonly used in the scientific domain that can be considered as MWEs due to the statistical criteria defined by Simpson-Vlach and Ellis (2010).

Moreover, we label each identified MWE with functional categories to assess (a) the functions MWEs have fulfilled in scientific writing across 300 years, and (b) whether there have been changes in their usage over time. We derive the functions *stance expressions, discourse organizers*, and *referential expressions* from extensive previous work based on Hallidayan register theory (Halliday and Matthiessen, 2014) and widely used by EAP researchers (Biber et al., 2004; Simpson-Vlach and Ellis, 2010; Liu, 2012). Finally, to assess change regarding MWEs, we employ a method from language modeling, relative entropy (Kullback-Leibler Divergence).

The remainder of the paper is organized as follows. In Section 2 we discuss related work on functional categories of MWEs. Sections 3 and 4 present our methods and results. We conclude with a summary of our findings and perspectives for future work (Section 5).

2. Related Work

There are numerous corpus-based accounts regarding the usage of MWEs in different registers, including the scientific one (e.g. Biber and Barbieri (2007); Hyland (2008); Liu (2012)), considering also their classification in terms of functions (see Biber et al. (2004); Simpson-Vlach and Ellis (2010) and Oakey (2020) for an overview). These studies are usually based on strategies for identifying formulaic, pre-fabricated, chunk-like and otherwise phraseological linguistic items considering frequency-based measures (such as MPI) derived from corpora (see work on lexical bundles (Biber and Barbieri, 2007; Hyland, 2008), academic formulas (Simpson-Vlach and Ellis, 2010), and multi-word constructions (Liu, 2012)).

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 81–87. May 25, 2024.

Computational linguistic accounts usually focus on techniques to identify and describe patterns of co-occurrence of linguistic units (e.g.Evert (2005); Gries (2022)). To identify potential MWE candidates different measures are applied. Gries (2022) proposes a strategy based on eight different dimensions of information, while Simpson-Vlach and Ellis (2010) define a formula teaching worth (FTW) score based on frequency and mutual information. The identification of MWEs using machine-learning methods are typically based either on DiMSUM (Schneider et al., 2016) or PARSEME (Savary et al., 2015) corpora and the complexity of this task can be attested by the low F1-scores of the state-of-the-art tools (i.e., below 65 as presented by Tanner and Hoffman (2023)). PARSEME corpus divides MWEs into different categories, but they are based on structural properties, not on their functions. Moreover, these datasets are not composed of scientific texts, and thus not totally suitable to address our research question.

Although the study of MWEs is a very active field, both from a linguistic and a computational point of view, the diachronic development of MWEs and their functions remains under-researched. While Biber et al. (2004) and Simpson-Vlach and Ellis (2010) propose a classification of MWEs in terms of discourse functions, these categories have not been examined diachronically. Alves et al. (2024) presented a study concerning the development of MWEs association metrics in scientific English, however, MWE functions were not the main focus of the analysis. Consequently, there are hardly any ready-to-use methodological approaches. With our work, we intend to fill these gaps.

3. Data and Methods

3.1. Data

As a data source, we use the Royal Society Corpus (RSC) 6.0[1], a diachronic corpus of scientific English covering the period from 1665 until 1996. This resource comprises 47,837 texts (295,895,749 tokens), mainly scientific articles covering a wide range of areas from mathematics to physical and biological sciences, and is based on the Philosophical Transactions and Proceedings of the Royal Society of London (Fischer et al., 2020). Table 1 shows a detailed overview of the distribution of texts and tokens over time.

Period	Texts	Tokens
1665–1699	1,325	2,582,856
1700–1749	1,686	3,414,795
1750–1799	1,819	6,342,489
1800–1849	2,774	9,112,274
1850–1899	6,754	36,993,412
1900–1949	10,011	65,431,384
1950–1996	23,468	172,018,539

Table 1: Size of the Royal Society Corpus 6.0 over time

There has been extensive work on the proceedings and transactions of the Royal Society based on the RSC, showing how the scientific register has evolved from an involved verbal style of writing (papers were read out aloud by fellows at the Royal Society of London in the beginning of the society) to a highly informational style of writing meant for purely written expert-to-expert communication. This development is specific to scientific writing and not observed in a register-mixed corpus (cf. Degaetano-Ortlieb and Teich (2019)). Also, we observe diversification in linguistic usage reflecting disciplinary specialization (e.g. modern chemistry emerges in the 18th c.) (Bizzoni et al., 2020) and a general conventionalization trend (Teich et al., 2021). Together, linguistic diversification and conventionalization address the communicative demands of modern science communication. In this paper, we expand this research by specifically investigating MWEs in the RSC since they are highly conventionalized structures.

3.2. Identifying MWEs in the RSC

In the present study, we focus on two specific kinds of MWEs that were extracted from the RSC using two different approaches: (a) fixed MWEs extracted from the UD-parsed version of the RSC, and (b) ensemble of MWEs provided by the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010).

Fixed Multi-word Expressions The Universal Dependencies[2] (UD) guidelines for morphosyntactic annotations (De Marneffe et al., 2021) encompass the relation label *fixed* for certain fixed grammaticalized expressions which tend to behave like function words (e.g. *because of*, *in spite of*, *as well as*) with distinct functions.

To extract the fixed MWEs, we parsed the RSC 6.0 using Stanza tool (Qi et al., 2020) with the combined model for the English language trained on different UD corpora (i.e., EWT, GUM, GUMReddit, PUD, and Pronouns). Using a Python script with the pyconll library[3], we identified and counted the fixed MWEs in the RSC texts per year.[4]

From the list of fixed MWEs, we identified the 100 most frequent ones and manually annotated

[1] https://fedora.clarin-d.uni-saarland.de/rsc_v6/

[2] https://universaldependencies.org/

[3] https://github.com/pyconll/pyconll

[4] A manual evaluation of 70 sentences (10 per 50-year period of the RSC) showed that the labelled attachment score of the parser is equal or higher than 85% for fixed MWEs in the different time periods.

Function	Type	MWEs	Examples
Stance	epistemic	84	*it is important*, *according to*
	attitudinal/modality	24	*we have to*, *needs to be*
	intention/prediction	11	*if you want to*, *to do so*
	ability	34	*can be found*, *it is possible to*
Discourse	topic introduction/focus	31	*in this article*, *for example in*
	topic elaboration/clarification	70	*due to the fact*, *the reason for*
Reference	identification/focus	61	*such as the*, *as can be seen in*
	imprecision	3	*and so on*, *and so forth*
	specification of attributes	177	*a form of*, *on the basis of*
	time/place/text reference	57	*at the end of*, *in between*

Table 2: Functional categories and types (cf. Biber et al. (2004)).

them according to the taxonomy in Section 3.2.[5] Since we consider only the fixed MWEs with high frequency in the RSC and conducted a manual evaluation of the identified expressions, we assume that the parsing errors have been minimized in this study.

AFL Multi-word Expressions The Academic Formulas List is an inventory of the most common formulaic sequences in academic English. It is composed of: a) a core list of 207 formulaic expressions found in written and spoken academic language (e.g.*in terms of* and *at the same time*; b) 200 expressions from written corpora (e.g. *on the other hand* and *it should be noted*); and c) 200 MWEs extracted from spoken academic English texts (e.g. *be able to* and *if you look at*) (Simpson-Vlach and Ellis, 2010). The AFL MWEs were identified by the authors with a special measure of usefulness called the formula teaching worth (FTW), which combines frequency and mutual information measures. Thus, the classification of the formulaic expressions from the AFL as MWEs is done due this statistical criterion.

3.3. MWE Functional Categories

We follow the taxonomy proposed by Biber et al. (2004), which captures the major functions of MWEs with three primary categories: (a) stance expressions, which express attitudes or assessments of certainty, framing other propositions; (b) discourse organizers that reflect relationships between parts of the discourse; and (c) referential expressions that refer to physical or abstract entities, or to the textual context, identifying a specific entity or pointing out to a specific attribute of it.

Table 2 presents a summarized version of the taxonomy established by Biber et al. (2004) (i.e., functions and types) together with the number of MWEs per type and examples observed in the RSC.

Note that Simpson-Vlach and Ellis (2010) classified most of the AFL MWEs according to a taxonomy similar to the one proposed by Biber et al. (2004). We selected these categorised MWEs to be examined in this study, adjusting the taxonomy according to Table 2.

3.4. Modeling Change with Relative Entropy

To analyse the diachronic development of the different MWE functional categories, first, we examined the relative frequency per year.

To detect evolutionary trends, we applied relative entropy, specifically Kullback-Leibler Divergence (KLD; Kullback and Leibler (1951)), a method for comparing probability distributions measuring the number of additional bits needed to encode a given data set A when a (non-optimal) model based on a data set B is used for a set of elements X. In our case, A and B correspond to sub-sets of the RSC (e.g. time slices) and X, i.e. the ensemble of MWEs of each function.

$$D_{KL}(A\|B) = \sum_{x \in X} A(x) \log \left(\frac{A(x)}{B(x)} \right) \qquad (1)$$

KLD provides an indication of the degree of divergence between corpora and identifies the features that are primarily associated with a difference.[6]

To detect periods of change using KLD given each functional category (stance, discourse, and reference), we adopt the methodology described in Degaetano-Ortlieb and Teich (2018).[7] Basically, we compare 20-year windows of past and present language use sliding with a 5-year gap over the time line (e.g. t1=1665-1685, t2=1691-1711). By plotting the divergence for each comparison on the time line, we can inspect peaks or troughs which

[5] The annotation was made by a linguistics student and verified by two specialists

[6] Discrepancies regarding vocabulary size are controlled by applying Jelinek-Mercer smoothing with lambda 0.05 (cf. Zhai and Lafferty (2004) and Fankhauser et al. (2014)).

[7] Degaetano-Ortlieb and Teich (2018) make the code available at: https://stefaniadegaetano.com/code/

indicates a change. A peak indicates that the divergence of that features increases, and is thus *typical* of the future 20 years in comparison to the past 20 years. In particular, we consider the pointwise KLD, i.e. the individual KLD of each feature (here: either functions or types), in order to determine a feature's rise or decrease in typicality.

4. Results

4.1. Frequency-based Trends

Figure 1 presents the evolution of each main functional category per year by relative frequency (i.e., MWEs occurrence/no. of tokens of each period).

In general, all three functions present an increasing tendency across time until the beginning of the twentieth century. The usage of referential expressions (black line) has a considerable increase in the second half of the eighteenth century. Moreover, from 1925 on, while both discourse (blue) and reference MWEs (red) present a decreasing tendency, the use of stance expressions seems to steadily increase even though these expressions remain relatively low in frequency.

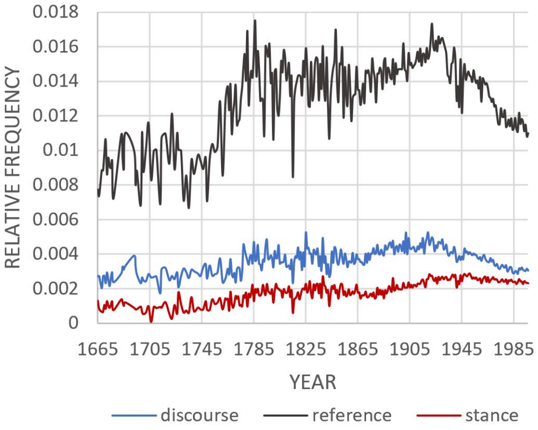


Figure 1: Relative frequency for each function.

4.2. Diachronic Trends by Divergence

While relative frequencies pinpoint the rise or decline of specific linguistic features over time, KLD provides a detailed quantification of the overall linguistic shift from one period to another, identifying even those changes that do not correspond to simple increases or decreases in usage frequency. Thus, KLD provides insights into the degree of linguistic change and allows to identify more subtle patterns of linguistic evolution that relative frequencies alone may not discern. Figure 2 presents the overall results per category for all the MWEs (AFL and fixed). We can observe that from the 17th to the beginning of 20th century, reference and discourse MWEs tend to behave in opposite directions, i.e. when reference becomes typical, discourse goes down in typicality and vice versa, while stance MWEs present less change. The scenario changes in the 20th century when the presence of stance expressions in the corpus becomes more typical.

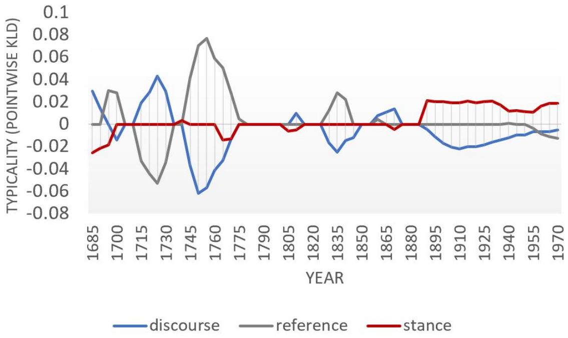


Figure 2: KLD measures for each function.

To better understand these diachronic trends, we also applied KLD considering the types of each function (see Figure 3). The main trends observed for discourse and referential expressions are due to the function types 'topic elaboration/clarification' and 'specification of attributes types', respectively. While the topic elaboration/clarification function is used to signal further explication providing a clearer understanding or additional information related to the topic being discussed (e.g. *in order to*, *as a result*, *the reason for*), the specification of attributes function type serves as a way to provide framing information (e.g. *the way which*, *the level of*, *these two*), i.e. essentially specifying or detailing characteristics, qualities, or attributes of a subject. These trends may be influenced by a variety of factors. Historical and cultural contexts that value explicit reasoning may lead to a preference for elaborate discourse, while changes in academic standards and expectations could necessitate a more precise specification of attributes. The rise of particular disciplines and interdisciplinary research, along with technological advancements that shape information dissemination, could also play significant roles.

Considering the increase in divergence for stance expression in the more contemporary period, we can observe that the peak is indicated by three out of four types for stance expressions. By 1825, ability becomes more typical showing an increased distinctive use (e.g. *can be used/found/expressed*), followed by attitudinal expressions until almost 100 years later where they decrease in divergence around the 1930s, when epistemic expressions (e.g. *according to, at least*) become typical. Around that period, also identification and focus reference expressions (e.g. *there has been, can be seen*) increase in typicality as well as topic and introduc-

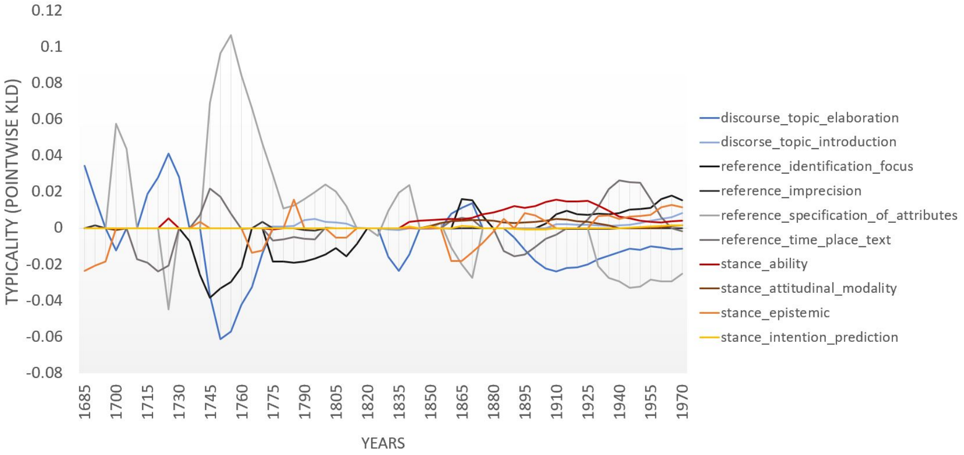


Figure 3: KLD measures for each function (colour) and its types (shades of a colour)

tion discourse organizers (e.g. *first of all, in this paper we*). During that period, there is also a peak in time, place and textual reference (e.g. *as shown in, shown in figure*). Overall, there is a trend towards a more varied distinctive use of MWE function types towards the more contemporary period. These trends seem to signal a use of MWEs to be increasingly inclined to articulate evidence-based reasoning as shown by MWEs such as *according to* or *as shown in*. These expressions serve to direct the reader's attention to evidence or examples that support the argument being made, which is a fundamental aspect of scholarly work.

5. Conclusion and Future Work

In this paper we have presented an analysis of MWEs in scientific writing, tracing the evolution of their functions over a span of three centuries. Our investigation reveals a dynamic landscape of MWE usage, marked by significant shifts in function that reflect changing priorities and practices within the scientific community over time. In the initial stages, we observed a competitive relationship between discourse and reference functions of MWEs. This competition underscores the evolving nature of scientific discourse, as authors sought to balance the needs for clarity and precision with the demands of argumentation and discourse structuring. Towards the recent 100 years, our findings indicate a diversification in MWE functions, with stance expressions taking on a leading role. The shift towards epistemic stance, reference of identification/focus, of place/time/textual and discourse organizers of topic and introduction seems to be a means of directing the reader's attention to evidence-based information.

Combining the AFL list with a UD-based approach to identify MWEs not covered by the AFL, allowed us to capture a broader range of conventionalized expressions that contribute to the diachronic trend of increasing conventionalization in scientific writing. The application of relative entropy as a methodological tool has further enriched our understanding of change over time, offering a quantitative measure of the shifts in MWE usage.

The functional categorization of MWEs, grounded in Hallidayan register theory, provides a solid theoretical framework for our analysis of functions and types. A limitation of our study is the uneven distribution of data across periods, with more material from recent periods, which may skew perceptions of MWE functionality and its evolution over time. Also, the diachrony of our data might present gaps within the AFL list. In future work, we aim to expand our research in three ways: (1) increase the number of MWEs related to the different functions and compare the obtained results with analysis of other domains; (2) model MWEs at the paradigmatic level by word embeddings to further increase coverage of items; (3) apply probabilistic measures of processing (e.g. surprisal) to gain insights on processing effects of conventionalization of MWEs. Overall, we aim to work towards gaining further insights into the complex ways MWEs serve the communicative needs of scientific writers and compare their usage across scientific domains and other registers.

Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

6. Bibliographical References

Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024. Multi-word expressions in english scientific writing. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 67–76, St. Julians, Malta. Association for Computational Linguistics.

Douglas Biber and Federica Barbieri. 2007. Lexical Bundles in University Spoken and Written Registers. *English for specific purposes*, 26(3):263–286.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at...: Lexical Bundles in University Teaching and Textbooks. *Applied linguistics*, 25(3):371–405.

Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. 2020. Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach. *Frontiers in Artificial Intelligence*, 3.

Kathy Conklin and Norbert Schmitt. 2012. The Processing of Formulaic Language. *Annual Review of Applied Linguistics*, 32:45–61.

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at COLING2018*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

Stefania Degaetano-Ortlieb and Elke Teich. 2019. Toward an Optimal Code for Communication: The Case of Scientific English. *Corpus Linguistics and Linguistic Theory*, 0(0):1–33. Ahead of print.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Universität Stuttgart.

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and Visualizing Variation in Language Resources. In *LREC*, pages 4125–4128.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 794–802.

Stefan Th. Gries. 2022. Multi-word Units (and Tokenization More Generally): a Multi-dimensional and Largely Information-theoretic Approach. *Lexis. Journal in English Lexicology*, (19).

MAK Halliday and CMIM Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*, volume 17. Routledge.

Ken Hyland. 2008. As can be seen: Lexical Bundles and Disciplinary Variation. *English for specific purposes*, 27(1):4–21.

Solomon Kullback and Richard A Leibler. 1951. On Information and Sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Dilin Liu. 2012. The Most Frequently-used Multi-word Constructions in Academic Written English: A Multi-corpus Study. *English for Specific Purposes*, 31(1):25–35.

David Oakey. 2020. Phrases in EAP Academic Writing Pedagogy: Illuminating Halliday's Influence on Research and Practice, journal = Journal of English for Academic Purposes. 44:100829.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European Multilingual Network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559.

Rita Simpson-Vlach and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied linguistics*, 31(4):487–512.

Joshua Tanner and Jacob Hoffman. 2023. MWE as WSD: Solving Multiword Expression Identification with Word Sense Disambiguation. *arXiv preprint arXiv:2303.06623*.

Elke Teich, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni. 2021. Less is More/More Diverse: On The Communicative Utility of Linguistic Conventionalization. *Frontiers in Communication*, 5.

Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.

Lexicons Gain the Upper Hand in Arabic MWE Identification

Najet Hadj Mohamed[12], **Agata Savary**[3], **Cherifa Ben Khelil**[14],
Jean-Yves Antoine[15], **Iskander Keskes**[2], **Lamia Belguith Hadrich**[2]

LIFAT - University of Tours[1]
MIRACL - University of Sfax [2]
LISN - University of Paris-Saclay [3]
EFREI Research Lab - University of Paris Panthéon Assas[4]
LIFO - University of Orleans[5]

najat.hadjmohamed@etu.univ-tours.fr, agata.savary@universite-paris-saclay.fr,
cherifa.ben-khelil@efrei.fr, jean-yves.antoine@univ-tours.fr, iskandarkeskes@fsegs.usf.tn,
lamia.belguith@fsegs.usf.tn

Abstract

This paper highlights the importance of integrating MWE identification with the development of syntactic MWE lexicons. It suggests that lexicons with minimal morphosyntactic information can amplify current MWE-annotated datasets and refine identification strategies. To our knowledge, this work represents the first attempt to focus on both seen and unseen of VMWEs for Arabic. It also deals with the challenge of differentiating between literal and figurative interpretations of idiomatic expressions. The approach involves a dual-phase procedure: first projecting a VMWE lexicon onto a corpus to identify candidate occurrences, then disambiguating these occurrences to distinguish idiomatic from literal instances. Experiments outlined in the paper aim to assess the efficacy of this technique, utilizing a lexicon known as LEXAR and the "parseme-ar" corpus. The findings suggest that lexicon-driven strategies have the potential to refine MWE identification, particularly for unseen occurrences.

Keywords: Multiword Expressions, Idiomatic Expressions, Literal vs. Figurative Meanings, Lexicon Augmentation, Arabic Language

1. Introduction

Multiword Expressions (MWEs) are a subject of interest across various fields related to language studies. They are part of each language's lexicon, distinct from literal words due to their non-compositional, preconstructed nature. Recently, the identification and analysis of MWEs have garnered significant attention in the field of Natural Language Processing (NLP), owing to their prevalence and nuanced semantic complexities. Despite considerable efforts in MWE identification, researchers have encountered challenges in addressing the issue of unseen MWE instances[1] (Taslimipoor et al., 2020; Pasquer et al., 2020b; Yirmibeşoğlu and Güngör, 2020; Kurfali, 2020). Savary et al. (2019) assert that to make substantial progress in MWE identification, it is imperative for the research community to integrate the identification process with the development of syntactic MWE lexicons. They advocate for lexicons that provide minimal morphosyntactic information, augmenting existing MWE-annotated corpora. This approach, they argue, complements traditional corpus-based methods with MWEs that occur rarely or never in MWE-annotated corpora. In this paper, we align ourselves with the same perspective, emphasizing the critical role of MWE lexicons in advancing MWE identification methodologies for Arabic language.

MWEs assume a unique and challenging role within this domain due to their non-compositionality and their ability to take on a figurative or literal meanings. For instance, the degree of transparency varies from one idiom to another. Thus, the following idiom is rather transparent سَلك الطَريق السَريع (salk al-ṭarīq al-sarīᶜ | lit. 'to take the fast road') 'to choose the easier way', i.e. it is easy to recover the motivation behind the image of taking a fast road. Conversely, in كَسَرَ السَيْف (kasara al-saif | lit. 'broke the sword') ' to triumph over an opponent or a difficult circumstance', the motivation for the image is unclear. Moreover, transparency can depend on the particular speaker's knowledge. For instance, the literal reading (e.g.'وَضَعَ يَداً عَلَى الجُرْح' (lit. 'to touch the wound') 'to evoke someone's weakness' is understandable for most speakers, while understanding the origin of the following idiom calls for historic and cultural knowledge: 'براءة الذئب من دم ابن يعقوب' ('brāᵓẗ al-ḏᵓib mn dm abn īᶜqūb' | lit. 'to have the innocence of the wolf from the Jacob's son blood') 'to be innocent'.[2]

[1] No other verbal multi-word expression containing the exact same set of lemmas has been annotated at least once in the training corpus.

[2] This idiom relates to the story of Jacob and his broth-

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 88–97. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

Significant research has been dedicated to detecting metaphors and understanding idiomatic expressions. Metaphors are deliberately constructed to convey figurative meanings, while idiomatic expressions can be interpreted either literally or figuratively, depending on the context of use (Shutova, 2010; Mason, 2004; Liu and Hwa, 2017). The accurate processing of idiomaticity within textual sequences is fundamental in NLP, given that idiomatic expressions constitute a significant aspect of linguistic communication. Attaining high performance in this task holds the potential to enhance various downstream applications, including sentiment analysis, information retrieval, and machine translation (Hashempour and Villavicencio, 2020; Mohamed et al., 2023). In this paper, our main focus is on identifying MWEs using an Arabic lexicon, with the goal of capturing unseen expressions more effectively and reducing the ambiguity of literal interpretations. Thus, we are also interested in the challenge of distinguishing between these two interpretations, which is complicated by the fact that idioms often do not follow easily identifiable linguistic patterns, especially for the Arabic language, given that is characterized by a fairly flexible word order (Hadj Mohamed et al., 2022). While our research primarily focuses on Arabic, we have also tested our model for the binary disambiguation of Potential Idiomatic Expression (PIE) task (see Section 2 on English and German languages. The paper is organized as follows: Section 2 provides a thorough review of existing literature on MWE identification. Section 3 focuses on MWE identification in Arabic. Following that, Section 4 elaborates on our methodology for MWE identification in Arabic, emphasizing the integration of lexicons and the disambiguation process, while Section 5 details the data used in our experiments. Finally, in Section 6, we present and analyze our experimental results.

2. Related work

A considerable amount of research has focused on MWE-specific tasks. In this paper we are primarily concerned with **MWE identification**, which consists in automatically annotating MWE occurrences in running text (Constant et al., 2017). Most approaches to this task are supervised, i.e. trained on manually annotated datasets, such as STREUSLE (Schneider and Smith, 2015) or PARSEME (Savary et al., 2018). Shared tasks such as DiMSUM (Schneider et al., 2016) and PARSEME (Ramisch et al., 2020) boosted the development of such tools. MWE identifiers are then trained and evaluated on these corpora. For instance, two approaches to MWE identification within a transition system were compared in (Al Saied et al., 2019): one based on a multilayer perceptron and the second on a linear SVM. Both approaches utilize only lemmas and morphosyntactic annotations from the corpus and were trained and tested on PARSEME Shared Task 1.1 data (Ramisch et al., 2018). The approach in (Kurfali, 2020) leverages feature-independent models with standard BERT embeddings. mBERT was also tested, but with lower results. An LSTM-CRF architecture combined with a rich set of features: word embedding, its POS tag, dependency relation, and its head word is proposed in (Yirmibeşoğlu and Güngör, 2020). The main focus of PARSEME Shared Task 1.2 was the detection of the unseen Verbal Multiword Expressions (VMWEs) which is more challenging compared to the identification of seen VMWEs (Ramisch et al., 2018). Several systems participated in the shared task, including MTLB-STRUCT (Taslimipoor et al., 2020), TRAVIS-mono and TRAVIS-multi developed by Kurfali (2020), Seen2Unseen developed by Pasquer et al. (2020a), ERMI by Yirmibeşoğlu and Güngör (2020) and others. Notably, the MTLB-STRUCT system, which leverages multilingual BERT fine-tuned for joint parsing and MWE identification, achieved the top cross-lingual macro-average in the open track for both the identification of VMWES and the subtask of identifying unseen VMWEs.

Since unseen VMWEs prove critically hard to identify, a natural idea would be to leverage the advances of **MWE discovery**, which consists finding new MWEs (types) in text corpora, and storing them for future use in a lexicon (Constant et al., 2017). Very many different approaches were devised for this task in the past, based on statistical association measures (Evert, 2005), parsing data (Seretan et al., 2011), lexico-syntactic constraints (Broda et al., 2008), possibly combined with the use of neural network (Pecina, 2010), etc.

An alternative approach in addressing unseen data, and the scarceness of MWE-annotated corpora in general, is to use existing **MWE lexicons**, extracted for instance from classical human-readable dictionaries (Kanclerz and Piasecki, 2022) or Wiktionary (Muzny and Zettlemoyer, 2013), possibly with example sentences contained therein (Tedeschi et al., 2022). Such a lexicon can be straightforwardly projected on a corpus by form/lemma matching. Each resulting word co-occurrence is then considered as a *potential idiomatic expression* (PIE), in the sense that it can be true idiomatic occurrence of a MWE, or just a literal/coincidental co-occurrence of the MWE component words.

The task of **binary disambiguation of PIEs** has been addressed by a number of works. Sporleder

ers, shared by the Jewish, Christian and Muslim religions.

and Li (2009) propose a generalized method utilizing cohesion graphs, hypothesizing that a PIE is used figuratively if its removal improves cohesion. Liu and Hwa (2018) introduce a ”literal usage metric” quantifying the literalness of a PIE, computed as the average similarity between words in the sentence and a literal usage representation. Ehren et al. used a 2-layer LSTM network to get latent representations for the verbal idiom tokens. These were then used in a fully connected layer to predict the class using softmax. They used pretrained static and contextualized word embeddings as an input for their model. In recent years, several shared tasks have been organized to advance research in binary PIE disambiguation. Notably, the Multilingual Idiomaticity Detection and Sentence Embedding shared task (Madabushi et al., 2022) has gained attention. It comprises two subtasks: (a) binary disambiguation of PIEs, and (b) semantic text similarity detection, including sentences with and without MWEs.

3. Arabic and MWEs processing

The ”Arabic language” includes Modern Standard Arabic (MSA) and diverse Arabic dialects. MSA is used in religious texts, poetry, and formal writing, while dialects are spoken in everyday conversation. In this section, we provide an overview of MSA’s distinctive characteristics and review previous research on the automatic processing of MWEs in Arabic, with a specific focus on MSA rather than dialectal forms.

In MSA, capitalization is absent, and the usage of punctuation marks is infrequent in contemporary Arabic texts. Additionally, this language commonly features long, complex sentences with right-to-left writing, often resulting in paragraphs that lack punctuation. Furthermore, as a Semitic language, Arabic exhibits a complex morphology. It uses *concatenative morphology (agglutinated or compound words)*, where words are formed via a sequential concatenation process[3]. For example, the sentence *‘then they will write it’* is presented in Arabic as one word فسيكتبونها. Moreover, Arabic includes words that can be altered with diacritical marks, either above or below them, creating new words with distinct pronunciations and meanings, often similar to the original word. Consequently, texts lacking diacritical marks are prone to ambiguity.

In Arabic, as in German, the word order is flexible, allowing specific words in a sentence to be rearranged without altering its meaning. This adaptability is achieved through the language’s use of case markers, particles, and other linguistic mechanisms to clarify word relationships, resulting in a more versatile syntax compared to languages with a more rigid word order. These unique features make Arabic a challenging language for NLP tasks.

Several studies and research have been conducted on Arabic Multiword Expressions (AMWEs). Attia (2006) explored AMWEs using a finite-state machinery and Lexical Functional Grammar (LFG). During processing, fixed and adjacent semi-fixed MWEs were scrutinized using lexical transducers, deconstructing one-word phrases into segments and integrating MWEs into spaced words. Syntactically flexible MWEs were handled by grammar rules as syntactically compositional but semantically non-compositional due to lexical selection rules. Attia et al. (2010) introduced a linguistic method based on regular expressions for extracting AMWEs from texts, with a specific focus on nominal AMWEs. Hawwari et al. (2014) compiled an AMWE list from 5,000 expressions extracted from dictionaries.(Al-Badrashiny et al., 2016) employed a paradigm detection method on the Arabic Treebank and Arabic Gigawords corpus, resulting in the autonomous extraction of 1,884 AMWEs, each displaying various forms due to morphological variations. Recently, as part of the PARSEME framework (Savary et al., 2023), Hadj Mohamed et al. (2022) manually constructed a corpus comprising 4,700 instances of Verbal AMWEs.

4. Method

Our ultimate goal is to address the task of identifying VMWEs in Arabic. However, within this paper, we specifically concentrate on the critical challenge of detecting unseen instances, which represents a significant frontier in the field. Our approach relies on a lexicon and minimizes noise by filtering out literal interpretations. In contrast to numerous existing methods for VMWE identification, we choose not to rely on a VMWE-annotated corpus, opting instead for a carefully curated VMWE list. This decision stems from the limited representation of MWEs with literal and figurative meanings in resources such as Arabic Wiktionary, leading us to manually extract VMWEs from an exhaustive paper dictionary. Given this VMWE lexicon, our methodology unfolds in two phases: the first is the identification of VMWE candidates, while the second involves the disambiguation of these candidate occurrences, as outlined by Algorithm (1). We start by aligning the VMWE lexicon with the test corpus to identify potential VMWE candidates within the text. This process involves comparing the lexicon entries with the content of the

[3] Agglutination is the process, common in Arabic, of adjoining clitics from simple word forms to create more complex forms.

test corpus in order to detect instances where VMWEs may occur. Then, we apply a binary PIE disambiguation method to distinguish between idiomatic and literal instances among these candidates. VMWEs are identified from idiomatic occurrences, while literal instances are retained for further analysis as supplementary data.

The following sections provide more detailed descriptions of these two phases.

Algorithm 1 : Procedure for extracting and filtering sentences containing MWEs from the corpus

```
procedure ExtractAndFilter($C$, $L$, $model$)
    $literal \leftarrow []$
    $idiomatic \leftarrow []$
    for $mwe \in L$ do
        for $sentence \in C$ do
            if $mwe$ occurs in $sentence$ then
                $class \leftarrow$ PIEC[4]$(mwe, sentence)$
                if $classification$ is "literal" then
                    $literal$.append$(sentence)$
                else
                    $idiomatic$.append$(sentence)$
                end if
            end if
        end for
    end for
    return $literal, idiomatic$
end procedure
```

4.1. Identifying VMWE candidates

During this phase, VMWE candidates are identified based on the lemmas associated with each MWE in the lexicon. The use of multisets allows for the identification of candidates in any order, regardless of the syntactic dependency between them. For example, consider the first VMWE seen in the lexicon (**L**) in Figure 1: وضع يده (ūḍᶜ īdh | lit. ' put hand+his') 'put one's hand'.

In sentences **(1)** and **(2)** from the *parseme-ar* corpus, the three lemmas "وضع" (*'to put'*), "يد" (*'hand'*), and "ه" (*'his'*) are present, resulting in their extraction as VMWE candidates. However, sentence **(2)** contains no VMWEs but rather a coincidental occurrence. In contrast, the candidate identified from sentence **(4)** represents a literal occurrence for the third VMWE طار غرابه (tar ġurab-h | lit. 'his crow flew off') 'to get old'" in **L**. The choice of using a forward step of filtering is a matter of balance between precision and recall. The expected noise present in the identification phase results in good recall (R= 0.79) but low precision (P=0.41). Addressing this challenge, the second filtering phase (4.2) aims to enhance precision. We achieve this through the implementation of subtask (A) of the SemEval shared task (Madabushi et al., 2022).

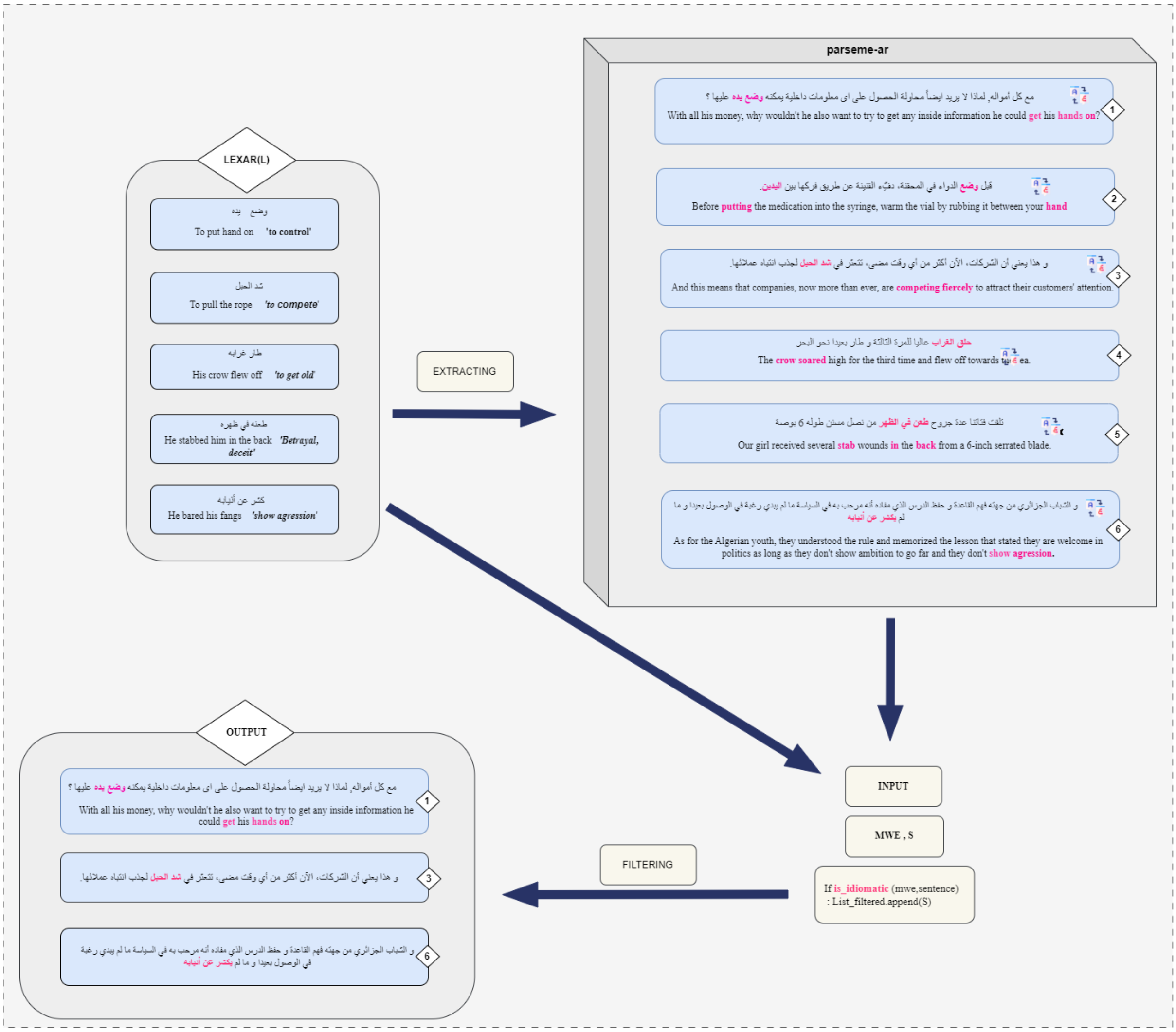
4.2. Disambiguating candidate VMWE occurrences

As previously stated, we proceed with our filtering phase by employing the same subtask (A) from the SemEval shared task. The aim here is to distinguish between the compositional (literal) and non-compositional (idiomatic) uses of PIE within a given context. This is different from the task of MWE extraction, which focuses on identifying MWEs within a corpus. Namely, our method takes a set of sentences containing a target PIE as input. We handle the disambiguation of PIEs in a manner similar to word sense disambiguation. Our fundamental assumption is that the context in which PIEs are used literally and figuratively differs significantly enough to justify distinct contextual representations. Figure 2 outlines an overview of the architecture, which is built upon the contextual language model used in our experiments, namely BERT.

Firstly, we aim to leverage the semantic idiosyncrasy characteristic of idiomatic expressions, highlighting that the meanings of the components within idiomatic expressions are related to the context in which they appear. To achieve this, we start by tokenizing the input, which consists of the sequence S and the target PIE. Following this, contextualized embeddings are generated using BERT and produce a vector representation for both the expression (PIE) and its context (S). Then, we add a Bidirectional LSTM (BiLSTM) layer for each embedding sequence to extract initial features from the raw embeddings. This results in in $h^{(S)}$ = BiLSTM$(e^{(S)})$ and $h^{(PIE)}$ = BiLSTM$(e^{(PIE)})$.

The attention flow layer integrates and combines information from both the context word sequence and the query word sequence (Seo et al., 2017). This process generates query-aware vector representations of the context words and propagates the word embeddings from the preceding layer. Similarly, in our specific task, the attention flow layer merges details from two embedding sequences that encode diverse types of information. We fused $h^{(S)}$ and $h^{(PIE)}$ into an attention layer to obtain an enhanced contextualized representations for both the sentence and the PIE. This results in a unified representation that integrates information from both the entire sentence and the PIE. Finally, we introduce a MaxPooling layer to reduce spatial dimensions in neural network architectures while preserving the most important features by selecting the maximum value from each feature map. Following this, the fused representation is passed through a series of Dense layers for classification.

The final output is produced by a sigmoid-



وضع يده
To put hand on 'to control'

شد الحبل
To pull the rope *'to compete'*

طار غرابه
His crow flew off *'to get old'*

طعنه في ظهره
He stabbed him in the back *'Betrayal, deceit'*

كشر عن أنيابه
He bared his fangs *'show agression'*

مع كل أمواله, لماذا لا يريد ايضاً محاولة الحصول على اى معلومات داخلية يمكنه وضع يده عليها ؟
With all his money, why wouldn't he also want to try to get any inside information he could get his hands on?

قبل وضع الدواء في المحقنة، دفّئ القنينة عن طريق فركها بين اليدين.
Before putting the medication into the syringe, warm the vial by rubbing it between your hand

و هذا يعني أن الشركات، الآن أكثر من أي وقت مضى، تتعثر في شد الحبل لجذب انتباه عملائها.
And this means that companies, now more than ever, are competing fiercely to attract their customers' attention.

حلق الغراب عاليا للمرة الثالثة و طار بعيدا نحو البحر
The crow soared high for the third time and flew off towards the sea.

تلقت فتاتنا عدة جروح طعن في الظهر من نصل مسنن طوله 6 بوصة
Our girl received several stab wounds in the back from a 6-inch serrated blade.

و الشباب الجزائري من جهته فهم القاعدة و حفظ الدرس الذي مفاده أنه مرحب به في السياسة ما لم يبدي رغبة في الوصول بعيدا و ما لم يكشر عن أنيابه
As for the Algerian youth, they understood the rule and memorized the lesson that stated they are welcome in politics as long as they don't show ambition to go far and they don't show agression.

If is_idiomatic (mwe,sentence) : List_filtered.append(S)

مع كل أمواله, لماذا لا يريد ايضاً محاولة الحصول على اى معلومات داخلية يمكنه وضع يده عليها ؟
With all his money, why wouldn't he also want to try to get any inside information he could get his hands on?

و هذا يعني أن الشركات، الآن أكثر من أي وقت مضى، تتعثر في شد الحبل لجذب انتباه عملائها.

و الشباب الجزائري من جهته فهم القاعدة و حفظ الدرس الذي مفاده أنه مرحب به في السياسة ما لم يبدي رغبة في الوصول بعيدا و ما لم يكشر عن أنيابه

Figure 1: Overview of the method.

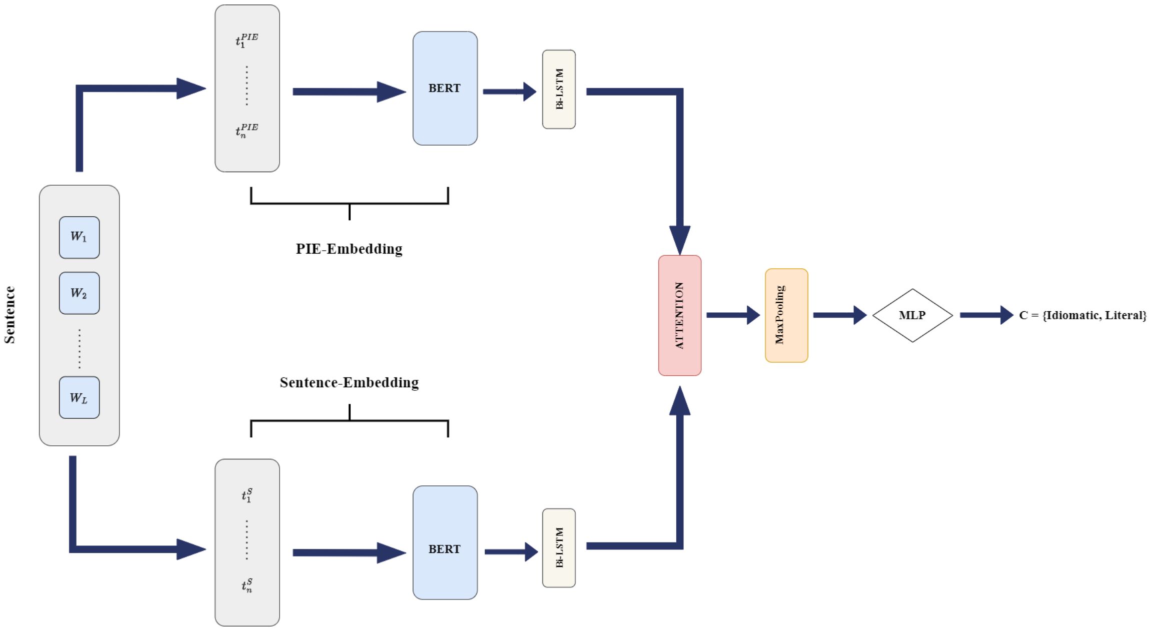


Figure 2: Overview of the PIEC model

activated Dense layer, providing a binary classification result (idiomatic or literal). Table 1 shows the hyper-parameters use with this architecture.

Parameter	Value
Sequence Length	128
Training Batch Size	256
Epoch number	30
Learning Rate	0.00001
Optimizer	Adam

Table 1: Model Training Parameters

5. Data

Assessing the efficacy of our MWE identification method necessitates both a VMWE lexicon and a corpus. As for the corpus, we used the ”parseme-ar” corpus from PARSEME 1.3 (Hadj Mohamed et al., 2022; Savary et al., 2023), which contains 4,7000 VMWEs within 7,500 sentences extracted from PADT belonging to the UD collection (Hajic et al., 2009). In our experiments, our focus was on two categories of VMWEs outlined in the parseme-ar corpus: LVC (Light Verb Construction) and VID (Verbal Idiom). We excluded the IAV (In Inherently Adpositional Verb) category, as it is considered optional. Following this, we manually created a lexicon named LEXAR[5], referenced as (**L**) in Figure 1. We meticulously extracted and compiled idiomatic expressions from ”Contextual Dictionary of Idiomatic Expressions” by Elsini (1998). Following the PARSEME annotation guidelines[6], we identified a total of 1504 Arabic VMWEs, and each expression in LEXAR underwent categorization by assigning a part-of-speech (POS) tag and determining its type as either LVC or VID. The annotation process, which took between 1-2 days and overlapped almost 70% of VMWEs with PARSEME-AR, ensured a comprehensive coverage of VMWEs in our corpus. We evaluated the performance of our idiomatic expression classifier, *PIEC*, by conducting evaluations with specialized datasets tailored to measure its accuracy in classifying sentences with idiomatic expressions. These evaluations encompassed datasets in Arabic, German, and English languages. Table 2 provides a summary of the data used to evaluate the secondary task. For Arabic, we trained the *PIEC* on a dataset included 34 idiomatic expressions. Each expression accompanied by sentences from the corpus of the shared task ConLL[7] encompassing both idiomatic and literal meanings. The 34 expressions were crafted manually by two native Arabic speakers. For instances lacking literal examples, we used ChatGPT to generate them, followed by manual verification. The MAGPIE corpus (Haagsma et al., 2020) provided the English dataset. It offers a collection of 1,756 PIEs, each representing different syntactic patterns, along with their associated sentences, totaling 56,622 annotated data instances with an average of 32.24 instances per PIE. For German we used the COLF-VID dataset (COrpus of Literal and Figurative meanings of Verbal IDioms) (Ehren et al., 2020). It contains 6,985 sentences sourced from newspaper articles, with annotations for 34 German VID types. Each MWE in the dataset is tagged with one of four labels: IDIOMATIC, LITERAL, UNDECIDABLE, or BOTH.

6. Results

The main goal of this study is to identify VMWEs, with a particular emphasis on unseen instances. Accordingly, we employed evaluation metrics aligned with the criteria of the shared task (Savary et al., 2017): These metrics include **MWE-based** metrics, which encompass precision, recall, and F1 scores for accurately detecting entire VMWEs, as well as precision, recall, and F1 measures for all VMWEs, including those that are unseen (**unseen MWE-based**). In Table 3, we compare the performance of our approach against MTLB-STRUCT.

On the multilingual level, MTLB-STRUCT achieved an MWE-based F1 score of 34.24 on unseen VMWEs and a global MWE-based F1 score of 56.27. Note that these results were obtained by re-training MTLB-STRUCT on the parseme-ar without the IAV category. However, even with the improvement in scores generated by the AraBert-based model (F1= 0.62 on the dev), Arabic is still one of the languages with the lowest performance score for global MWE-based and unseen-based scores. Although the F1 scores for unseen MWEs are still not optimal, our approach outperforms MTLB-STRUCT in terms of MWE-based F1 score by 7% and for unseen MWEs by 9%. Among the 278 unseen VMWEs assessed, our approach detected 125, whereas MTLB-STRUCT identified 104 out of the total.

For our experiments on the **binary disambiguation of PIEs** task (Figure 2), we focused only on the IDIOMATIC and LITERAL labels. Table 4 presents the results of our experiments on the TEST set. As baseline, we used a conventional SVM (Support Vector Machine) with MUSE (Multilingual Unsupervised and Supervised Embeddings) (Conneau et al., 2018) features. Em-

[5] We plan to release the lexicon upon acceptance of this paper

[6] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2

[7] https://lindat.mff.cuni.cz/

Lang	Literal	Figurative	Total
AR-train	103	202	305
AR-dev	16	30	46
AR-test	29	57	86
COLF-VID-train	1,172	5,705	6,902
COLF-VID-dev	264	1,214	1,488
COLF-VID-test	265	1,238	1,511
MAGPIE-train	2,676	12,676	15,352
MAGPIE-dev	595	2719	3314
MAGPIE-test	635	3339	3974

Table 2: Literal and idiomatic occurrences of PIEs in Arabic (AR), German (DE) (we excluded both the types of BOTH and UNDECIDABLE, which accounts for the disparity in the count between literal and idiomatic expressions compared to the total) and English(EN)

beddings were independently generated for both the PIE instances and sentences using the MUSE library. Notably, PIEC demonstrates better performance compared to the baseline MUSE-SVM. Including semantic information regarding both the context and the PIE significantly enhances the classifier's performance. It performs highly better on both literal and figurative class across all languages, even when dealing with unbalanced data in German and English. For instance, in the literal class, the F-score exhibited significant improvements: in Arabic from 0.44% to 0.89%, in English from 0.39% to 0.86%, and in German from 0.54% to 0.78%. Hence, the consistency of the PIEC classifier's performance with BERT embeddings implies that accurate disambiguation of PIEs across numerous languages can be achieved with good precision, necessitating only a small set of annotated sentences.

7. Conclusion

This paper introduces a simple yet impactful strategy for improving the identification of VMWE through the integration of lexicons, with our lexicon named LEXAR. Specifically focusing on the Arabic language, we demonstrate that our approach outperformed neural architectures like MTLB-STRUCT. Additionally, our method effectively adresses the challenge of binary disambiguation by employing contextual embeddings, which differentiate between various uses of the same lexical units and assign appropriate representations. Although detecting unseen MWEs proves to be a challenging task in our experiments, we achieve promising results using lexicons, surpassing the previous state-of-the-art. Moreover, our proposed model for the **binary disambiguation of PIEs** task shows significant potential for extension to multiple languages, facilitated by multilingual contextual embeddings.

Acknowledgement

We would like to express our sincere gratitude to Rafael Ehren and Laura Kallmeyer for graciously accepting me (Najet Hadj Mohamed) to undertake a short-term scientific mission at Heinrich Heine University Düsseldorf. We especially thank Rafael Ehren for providing the preprocessed English data, which significantly contributed to the completion of this research. Additionally, we extend our appreciation to UniDive, the CA21167 COST Action[8]: Universality, diversity and idiosyncrasy in language technology for their support and funding, which facilitated this study.

8. Bibliographical References

Mohamed Al-Badrashiny, Abdelati Hawwari, Mahmoud Ghoneim, and Mona Diab. 2016. SAMER: a semi-automatically created lexical resource for Arabic verbal multiword expressions tokens paradigm and their morphosyntactic features. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 113–122.

Hazem Al Saied, Marie Candito, and Mathieu Constant. 2019. Comparing linear and neural models for competitive mwe identification. In *The 22nd Nordic Conference on Computational Linguistics*.

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, and Josef Van Genabith. 2010. Automatic extraction of Arabic multiword expressions. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 19–27.

Mohammed A. Attia. 2006. Accommodating multiword expressions in an Arabic LFG grammar. In

[8] https://www.cost.eu/actions/CA21167/

Lang	Our approach						MTLB-STRUCT					
	MWE-based			*unseen MWE-based*			*MWE-based*			*unseen MWE-based*		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Arabic	64.87	61.91	63.36	44.88	41.67	43.21	55.07	57.35	56.27	37.77	31.47	34.24

Table 3: Comparing our approach performance with MTLB-STRUCT on MWE-based and unseen MWE-based metrics.

Lang	SVM-MUSE						PIEC					
	Literal			*Figurative*			*Literal*			*Figurative*		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Arabic	0.40	0.50	0.44	0.82	0.75	0.78	0.90	0.88	0.89	0.96	0.96	0.96
English	0.81	0.26	0.39	0.84	0.98	0.91	0.92	0.82	0.86	0.96	0.98	0.97
German	0.79	0.41	0.54	0.89	0.98	0.93	0.80	0.77	0.78	0.95	0.96	0.95

Table 4: Comparing SVM-MUSE and PIEC performance across 3 languages in term of Precision (P), Recall (R), and F-measure (F1).

International Conference on Natural Language Processing (in Finland), pages 87–98. Springer.

Bartosz Broda, Maciej Piasecki, and Stanislaw Szpakowicz. 2008. Sense-based clustering of polish nouns in the extraction of semantic relatedness. In *2008 International Multiconference on Computer Science and Information Technology*, pages 83–89. IEEE.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of german verbal idioms with a bilstm architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 211–220.

Mahmoud Ismail Elsini. 1998. *Contextual dictionary of idiomatic expressions*. Lebanon Library Publishers.

Stefan Evert. 2005. *The statistics of word cooccurrences*. Ph.D. thesis, Dissertation, Stuttgart University.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287.

Najet Hadj Mohamed, Cherifa Ben Khelil, Agata Savary, Iskandar Keskes, Jean-Yves Antoine, and Lamia Belguith Hadrich. 2022. Annotating Verbal Multiword Expressions in Arabic: Assessing the Validity of a Multilingual Annotation Procedure. In *13th Conference on Language Resources and Evaluation (LREC 2022)*, Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1839–1848, Marseille, France.

Jan Hajic, Otakar Smrz, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. 2009. Prague arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, volume 1.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80.

Abdelati Hawwari, Mohammed Attia, and Mona Diab. 2014. A framework for the classification and annotation of multiword expressions in dialectal Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 48–56, Doha, Qatar. Association for Computational Linguistics.

Kamil Kanclerz and Maciej Piasecki. 2022. Deep neural representations for multiword expressions detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 444–453, Dublin, Ireland. Association for Computational Linguistics.

Murathan Kurfali. 2020. Travis at parseme shared task 2020: How good is (m) bert at seeing the unseen? In *International Conference on Computational Linguistics (COLING), Barcelona, Spain (Online), December 13, 2020*, pages 136–141.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1731.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. Semeval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. *arXiv preprint arXiv:2204.10050*.

Zachary J Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational linguistics*, 30(1):23–44.

Najet Hadj Mohamed, Malak Rassem, Lifeng Han, and Goran Nenadic. 2023. Alphamwe-arabic: Arabic edition of multilingual parallel corpora with multiword expression annotations. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 448–457.

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA. Association for Computational Linguistics.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020a. Seen2unseen at parseme shared task 2020: All roads do not lead to unseen verb-noun vmwes. In *Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LZX 2020)*.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020b. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language resources and evaluation*, 44:137–158.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten van Gompel, et al. 2018. Parseme multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Agata Savary, Silvio Ricardo Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91. Association for Computational Linguistics.

Agata Savary, Chérifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, et al. 2023. Parseme corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35.

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemi Zadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, et al.

2017. The parseme shared task on automatic identification of verbal multiword expressions. In *The 13th Workshop on Multiword Expression at EACL*, pages 31–47.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension iclr. *arXiv preprint arXiv:1611.01603*.

Violeta Seretan et al. 2011. *Syntax-based collocation extraction*, volume 44. Springer Dordrecht.

Ekaterina Shutova. 2010. Models of metaphor in nlp. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 688–697.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. Mtlb-struct@ parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. *arXiv preprint arXiv:2011.02541*.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Zeynep Yirmibeşoğlu and Tunga Güngör. 2020. Ermi at parseme shared task 2020: Embedding-rich multiword expression identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 130–135.

Revisiting VMWEs in Hindi: Annotating Layers of Predication

Kanishka Jain, Ashwini Vaidya
Department of Humanities and Social Sciences
Indian Institute of Technology, Delhi
{huz218481, avaidya}@iitd.ac.in

Abstract

Multiword expressions in languages like Hindi are both productive and challenging. Hindi not only uses a variety of verbal multiword expressions (VMWEs) but also employs different combinatorial strategies to create new types of multiword expressions. In this paper we are investigating two such strategies that are quite common in the language. Firstly, we describe that VMWEs in Hindi are not just lexical but also morphological. Causatives are formed morphologically in Hindi. Second, we examine Stacked VMWEs i.e. when at least two VMWEs occur together. We suggest that the existing PARSEME annotation framework can be extended to these two phenomena without changing the existing guidelines. We also propose rule-based heuristics using existing Universal Dependency annotations to automatically identify and annotate some of the VMWEs in the language. The goal of this paper is to refine the existing PARSEME corpus of Hindi for VMWEs while expanding its scope giving a more comprehensive picture of VMWEs in Hindi.

Keywords: Annotation, Stacked VMWE, Morphological Causative

1. Introduction

Verbal multiword expressions are linguistic constructions that involve multiple verbs or a combination of verb and other lexical item(s). These expressions combine to form new meanings (Baldwin and Kim, 2010). However, the non-compositional nature of these multiword expressions pose a challenge to any kind of natural language processing (NLP) task. Therefore, they have been part of multiple annotation efforts across languages.

The PARSEME shared task (Ramisch et al., 2020, 2018) is one such effort that aims to identify and annotate different types of VMWEs in multiple languages. We examine the Hindi corpus from the PARSEME shared task (Ramisch et al., 2020). In this paper, we have conducted a detailed survey of the corpus and identified some problems. A prominent issue that was prevalent across all annotation categories was missing annotations for a number of expressions. Another repeated issue that we observed is the annotation of modal constructions as multi-verbal constructions (MVCs) as both are structurally similar to each other. We address these and other issues in the existing corpus and refine the annotations to create a better quality dataset[1].

Multiword expressions in some languages are highly frequent. Hindi, for instance, in comparison to languages like English, is known to have a greater proportion of VMWEs compared to simple verbs (Vaidya et al., 2016). This productive usage of multiword expressions in the language has been captured in the PARSEME corpus edition 1.3 (Savary et al., 2023). But two additional and quite common phenomena need to be addressed. In Hindi, verbal complex allows for recursive combinations of light verb, multi-verb, and causative verbs. Sometimes all three can combine together. When two VMWEs appear together to create a single predicate then we refer to such predicate as Stacked VMWE. Further, VMWEs in Hindi are formed not only lexically (i.e. combining two or more lexical items) but also morphologically (i.e. combining two or more morphemes). In Hindi, morphological VMWEs occur as Causatives. Both, stacked and causative VMWEs are extensively used in the language but have not been explicitly annotated as such within existing annotation frameworks of the language.

The aim of this paper is twofold. First, to refine the existing corpus by addressing various issues and second, to extend its scope.

The paper is organized as follows. In Section 2 we describe different types of VMWEs found in Hindi. We also describe causatives and stacked VMWEs. Section 3 discusses the issues found in the annotations and how they have been addressed in the present study. Results and conclusion are presented in Section 4.

2. VMWEs in Hindi

2.1. PARSEME VMWEs

The PARSEME framework (Ramisch et al., 2020, 2018) has five categories of verbal multiword expressions (VMWEs) out of which three are tagged for Hindi i.e. Light Verb construction (LVC) as LVC.full and LVC.cause, Multi-Verb Construction (MVC), and Verbal Idiom (VID). The fundamental

[1] The dataset can be accessed from `https://gitlab.com/kjain93/mwe_ud_hindi`

difference among these categories lie in terms of their predication strategy. A VID has at least two elements combining – a main verb and its dependent which is not restricted to any one particular lexical category as shown in (1). On the contrary, LVC and MVC are formed with a preverbal element and a light verb. The only difference between the two categories is that the preverbal element in an LVC is noun whereas in case of an MVC it is a verb as shown in (2) and (3), respectively.

(1) bəɽti mehengai pər **ləgam**
increasing.F price-hike.F on rein.SG.F
ləgana zəruri hɛ
put.INF important.F be.PRS
'It is important to control the price-hike (or inflation).'

(2) ləɽke-ne gehnõ-ki
boy.3.SG.M-ERG jewellery.PL.M-GEN.F
cori ki
theft.F do.PST.F
'The boy has stolen the jewellery.'

(3) ləɽke-ne kɪtab **pəɽ^h**
boy.3.SG.M-ERG book.SG.F read
li
take.PST.SG.F
'The boy read the book (completely).'

Further, as mentioned above LVCs have been distinguished as LVC.full and LVC.cause. The difference is made in terms of the type of light verb used. If the light verb is 'causative' such that the subject is the cause of an event then it has been annotated as LVC.cause else as a LVC.full. An example is shown in (4). Compare it with its non-causative counterpart in (2). The subject /ləɽka/ 'boy' is the cause of an event of theft in (4) but an agent in (2). The causative meaning is expressed by the /-va/ morpheme on the verb in (4).

(4) ləɽke-ne naukar-se
boy.3.SG.M-ERG servant.3.SG.M-INST
gehnõ-ki **cori**
jewellery.PL.M-GEN.F theft.F
kər-va-yi
do-ICAUS-PST.PERF.SG.F
'The boy made the servant steal the jewellery.'

In the existing PARSEME corpus of Hindi a total of 1034 VMWEs have been annotated out of 35430 tokens as shown in Table 1. Further, it is to be noted that the frequency of VMWEs when compared to other Indo-European languages is quite high. These number are compiled from PARSEME shared tasks 2020[2] and 2018[3].

While the existing PARSEME framework covers all the prominent categories of VMWEs in Hindi, there are additional phenomena that are not present. The rest of the paper discusses two such phenomena – stacked VMWEs and causatives.

2.2. Morphological Causative

Causatives are common across natural languages. This is especially true for South-Asian languages like Hindi where any verb, theoretically, can undergo the morphological process and form causative. For instance, in (5b) the causative marker /-va/ attaches to the transitive verb /bənana/ 'build' and forms causative /bənvana/. The causativization of the transitive verb in (5a) increases the valency from two to three.

(5) a. ləɽke-ne ghər
boy.3.SG.M-ERG house-3.M
bənaya
build.PST.PERF.SG.M
'The boy built a house.'

b. ləɽke-ne bəcci-se
boy.3.SG.M-ERG girl.3.SG.F-INST
ghar
house.3.M
bən-va-ya
build-ICAUS-PST.PERF.SG.M
'The boy made the girl build the house.'

Apart from causativizing a simple verb, the language also allows causativization of light verbs[4] as shown in (4) where the light verb /ki/ 'do' is a causative.

Valency change is a property that is common to LVCs, MVCs and morphological causatives (Butt and King, 2006; Butt et al., 2008; Butt, 2010). For instance in (6a) simple verb /katna/ 'cut' has two argument positions – the servant and the tree. But in (6b) when katna/ combines with the light verb /dena/ 'give', forming an MVC, it has three argument positions. The new argument position for /laɽka/ 'boy' is licensed by the light verb /dena/ (Butt, 2010).

[2] http://multiword.sourceforge.net/mwelex2020

[3] http://multiword.sourceforge.net/lawmwecxg2018

[4] According to (Butt et al., 2008), Hindi also allows for causatives in MVC construction but we did not find examples of this in the current corpus

Langage	Tokens	VID	LVC.full	LVC.cause	MVC	Others	Total
English	124203	139	244	43	4	402	832
French	525992	2156	1878	97	22	1501	5654
German	173562	1437	311	33	0	2260	4041
Hindi	35430	61	641	26	306	0	1034
Italian	430789	1484	734	174	33	1785	4210

Table 1: Number of VMWEs in different Indo-European languages including Hindi in PARSEME shared tasks.

(6) a. naukər-ne paud^{h}a
servant.3.SG.M-ERG plant.SG.M
kata
cut.PST.PL.M
'The servant cut the plant.'

b. laɽke-ne naukər-ko
boy.3.SG.M-ERG servant.3.SG.M-DAT
paud^{h}a **katne**
plant.SG.M cut.INF.SG
dɪ-ya
give-PST.PERF.SG.M
'The boy let the servant cut the plant.'

This valency change is similar to causatives in example (5) where /-va/ morpheme combines with verb and license a new argument position for the causer 'girl'. This provides evidence that morphological VMWEs are similar to lexical VMWEs in Hindi. Hence, we propose to include them in the PARSEME framework.

PARSEME's existing annotation schema already annotates example like (4) as LVC.cause distinguishing them from their non-causative counterpart as in example (2) annotated as LVC.full. The addition of other causatives will then give a comprehensive picture of VMWEs in this language.

The examples discussed so far captures only one kind of causatives i.e. a causative formed by attaching /-va/ morpheme. They are also known as 'indirect causatives'. However, Hindi also has direct causatives that are formed by causativization of intransitive verbs as exemplified in (7).

(7) a. ləkɽi **jəli**
wood.SG.F burn.PST.F
'Wood burnt.'

b. ləɽke-ne ləkɽi
boy.3.SG.M-ERG wood.SG.F
jəl-a-yi
burn-DCAUS-PST.PERF.SG.F
'The boy burnt the wood.'

In (7a), the verb /jəlna/ 'burn' in intransitive whereas in (7b) the direct causative marker /-a/ is attached to the verb and forms the causative /jəlana/. Direct causatives, similar to indirect causatives, change the valency of the base verb from single argument place to two argument places. Therefore, direct causatives are also an example of morphologically formed multiword expressions.

In Hindi, direct causatives for some verbs are realized by a change in the phonological realization of the root of the verb as in (8) where the verb /d^{h}ul/ 'wash' changes to causative /d^{h}o/.

(8) a. kəpɽe **d^{h}ule**
cloth.PL wash.PERF.PL.M
'Clothes are washed.'

b. ləɽke-ne kəpɽe
boy.3.SG.M-ERG cloth.PL.M
d^{h}o-ye
wash.DCAUS-PERF.PL.M
'The boy has washed the clothes.'

These examples show that the system of morphological predication in the language is quite robust and complex. It is, therefore, essential to capture these various kinds of morphological multiword expressions to understand the representation of different types of VMWEs in Hindi. Hence, in this work we propose to annotate causatives using a morphological feature 'Cause' on verbs (see Section 3). The feature 'Cause' can effectively differentiates between the causative and non-causative forms of the verbs.

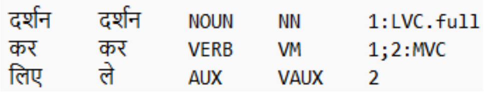
2.3. Recursive VMWEs

VMWEs in Hindi are not limited to combining two lexical items or morphological items but due to their recursive nature allow two or more VMWEs to stack describing a single event (Butt et al., 2003). An example is shown in (9) where an MVC is stacked on an LVC and results in a Stacked VMWE.

(9) ləɽke-ne gehnõ-ki
boy.3.SG.M-ERG jewellery.PL.M-GEN.F
cori kər dali
theft.F do put.PST.F
'The boy has stolen away the jewellery.'

In (9) there are three elements unlike the common pattern observed in LVCs and MVCs of predicating two elements. There is a noun /cori/ 'theft' and two verbs kər 'do' as well as /dali/ 'put'. The first or main verb can be in its base form or infinitive form whereas the second light verb is inflected for tense and aspect similar to MVC in the language.

Forming stacked VMWEs via recursion has not been implemented in an annotated corpus. Although PARSEME Hindi Corpus edition 1.3 does capture some of the stacked VMWEs as illustrated in Figure 1, it has not been discussed explicitly.



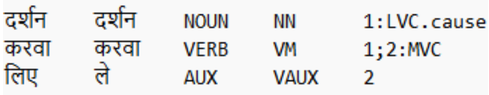
दर्शन	दर्शन	NOUN	NN	1:LVC.full
कर	कर	VERB	VM	1;2:MVC
लिए	ले	AUX	VAUX	2

Figure 1: An eaxample of LVC and MVC Stacked VMWEs in PARSEME Hindi Corpus edition 1.3. The noun /dərʃan/ 'sight' combines with the verb /kərna/ 'do' and a light verb /lena/ 'take'.

Further, recursivity in VMWEs can be seen at various levels thus resulting in layers of predication. In our example of LVC.cause in (4), the causative is stacked with an LVC forming an LVC.cause which can be further predicated with an MVC. The stacked VMWE in (10) thus shows stacking of three VMWEs – LVC+causative+MVC.

(10) ləɽke-ne naukar-se
boy.3.SG.M-ERG servant.3.SG.M-INST
gehnõ-ki **cori**
jewellery.PL.M-GEN.F theft.F
kər-va **dali**
do-ICAUS.SG.M put.PST.F
'The boy had the servant steal away the jewellery.'

The annotation of these layers of predication is shown in Figure (2).



दर्शन	दर्शन	NOUN	NN	1:LVC.cause
करवा	करवा	VERB	VM	1;2:MVC
लिए	ले	AUX	VAUX	2

Figure 2: An eaxample of LVC, Causative, and MVC Stacked VMWEs in PARSEME Hindi Corpus edition 1.3. The noun /dərʃan/ 'sight' combines with the verb /kərna/ 'do', indirect causative marker /va/, and a light verb /lena/ 'take'.

While VMWEs are formed via recursivity of existing multiword expressions, we do not intend to annotate them with a new label. Rather, we extract them using existing annotations which will be more efficient (see Section 3.2.3).

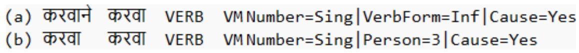
3. Enhancing the Annotations

The task of identifying multiword expressions is challenging and requires linguistic expertise. While the annotation guidelines developed as part of PARSEME shared task (Ramisch et al., 2020, 2018) standardizes the process of identification of VMWEs for many languages but there still exist various problems. In the following sections, we discuss some of the issues found in the PARSEME Hindi corpus edition 1.3 pertaining to existing annotation of VMWEs in Hindi and their refinement. We also discuss the annotations of morphological feature for causatives (Section 3.1) and representation of Stacked VMWEs (Section 3.2.3) in the existing annotation schema.

The PARSEME corpus of Hindi uses a treebank which is annotated using UD framework and therefore we could employ annotations for morphological description of tokens for automatic tagging of VMWEs.

3.1. Semi-Automated Annotation of morphological VMWEs

Beginning with causatives, we propose to add them as a morphological feature. If a verb is present in its causative form then we add 'Cause=Yes' as a boolean feature as illustrated in Figure (3). We note that Universal Dependencies guidelines have a similar feature 'Voice=Cau'[5]. In a future version of our corpus, we plan to update this feature to be in accordance with UD guidelines.



(a)	करवाने	करवा	VERB	VM	Number=Sing|VerbForm=Inf|Cause=Yes
(b)	करवा	करवा	VERB	VM	Number=Sing|Person=3|Cause=Yes

Figure 3: Feature structure for Hindi causative verb inflected for agreement /kərvane/ in (a) and /kərva/ in (b) with the 'cause' morphological feature. Note that the lemma form form for both the verbs is /kərva/.

The annotation process of causative verbs is semi-automatic as indirect causatives and one type of direct causative can be tagged using rule-based heuristics. The lemma form for /-va/ causatives have /-va/ attached however there are some discrepancies in the data therefore we have used a list of morphological endings with /-va/ morpheme varying only in terms of agreement features on the tokens to retrieve all indirect causative verbs.

The annotation of direct causatives was also challenging. Beginning with the /-a/ causatives, the UD framework does identify these causatives in their lemma. However, there are two issues in using them. First, as noted in case of indirect causative

[5] https://universaldependencies.org/u/feat/Voice.html

there are some inconsistencies with the identification of lemmas in the data. Second, Hindi also have other verbs ending with vowel /a/ like /ja/ 'go', /la/ 'get', and so on that are not causatives. Hence using only lemma leads to over-generation of tokens and to avoid that we have used multiple heuristics and manual checks while annotating the /-a/ causatives.

The second issue was with other type of direct causatives (c.f. example 8) where causative formation affects the phonological realization of the root and we get irregular forms. Since there is no particular pattern which can be exploited to identify these kind of verbs we have annotated them manually. A total of 269 causatives have been annotated – 165 automatically and 104 manually.

3.2. Automated Annotations of lexical VMWEs

Annotation of LVCs and MVCs was done in two stages, that is, automatic annotation using python scripts followed by manual adjudication. After annotating LVCs and MVCs we have extracted Stacked VMWEs.

3.2.1. LVCs

In this work, we aim to comprehensively annotate all the occurrences of VMWEs in the corpus. While examining the PARSEME corpus we observed that despite passing tests from the PARSEME guidelines a number of MVWEs were not annotated. Though it was true for all the categories, it was especially seen in case of LVCs (see Table 2 for comparision). Therefore, we used the dependency relation to find all the instances of LVCs in the corpus. Particularly, the 'compound' dependency relation that already identifies these noun+verb pairs have been used as in Figure 4.

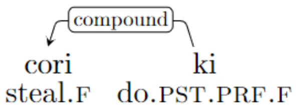


Figure 4: Compound dependency relation as tagged in UD framework for LVCs

All the missing LVCs were added to the existing corpus according to PARSEME guidelines. In order to distinguish between LVC.full and LVC.cause we use feature 'cause', annotated previously. For the purpose of this work, we have limited LVC.cause to only indirect caustives and have not included direct causatives.

We have also manually adjudicated the corpus using PARSEME tests for LVCs to remove any erroneous cases that have been annotated. Since, automatic annotations were dependent on UD dependency relation, we found few instances where nouns that were not abstract have been identified to be in compound relation with a verb as shown in (11)

(11) d^{h}ən lɪ-ya
money.M take-PST.PERF.M
'took money'

In (11), /d^{h}ən/ 'money' is annotated for compound relation with verb lɪya 'take'. These were not annotated as LVCs.

Data	LVC full	LVC cause
PARSEME	641	26
New	743	40

Table 2: Number of LVCs in existing PARSEME corpus and the new corpus.

3.2.2. MVCs

MVCs as discussed in Section 1 are formed by the combination of verb with a light verb. However, this pattern is confusable with other types of constructions in Hindi. For instance, both modal and passive constructions are superficially similar to MVCs. Modal verbs include examples like /pa 'able', and sək/ 'can/may' (example 12). /pa/ is ambiguous such that the same form occurs both as a simple verb meaning 'to get' and as a ability modal (Bhatt et al., 2011). As a simple verb, it can form a complex predicate and occur as a preverbal but it does not occur as a light verb. The current guidelines of PARSEME includes it as a light verb, however according to our current analysis the guidelines for Hindi needs to be updated to prevent confusion with modals.

For both MVCs and modals, the main verb appears in its base form while light verbs and modals are inflected for agreement features (Butt and Ramchand, 2005), as shown in (12).

(12) ləɽka kɪtab **pəɽ^h**
boy.3.SG.M book.SG.F read
pa-ya
can-PST.PERF.SG.M
'The boy could read the book.'

Constructions like (12) will pass the PARSEME tests for tagging MVCs, however, semantically there is a difference between light verbs and modals. Light verbs contribute sub-event information as seen in (13), where light verb /dɪya/ 'give' contributes permissive meaning to the event (Butt, 1995). Modals, on the other hand, place an event into possible world semantics (Butt, 2010) (example (12)).

(13) ləɽke-ne naukar-ko
boy.3.SG.M-ERG servant.3.SG.M-DAT
xət **pəɽʰne dɪ-ya**
letter.SG.M read.INF give-PST.PERF.SG.M
'The boy let the servant read the letter.'

Similarly, verbs in passive constructions appear by combining any main verb with an auxiliary verb /ja/ 'go' as shown in (14). The /ja/ 'go' can participate in a number of constructions. It can be used as a simple verb with the meaning ' to go', as a light verb with the meaning 'with force' and also as an auxiliary when a sentence is passivized. On the surface, the passive resembles MVCs where two verbs are predicated and are incorrectly annotated as MVCs in the current PARSEME corpus of Hindi at several places.

(14) ləɽke-se kɪtab
boy.3.SG.M-INST book.SG.F
pəɽʰi gə-yi
read.PST.SG.F go-PST.PERF.SG.F
'The book was read by the boy.'

The main verb in passives, for example pəɽʰi 'read', in (14), is inflected for tense and aspect which violates the first test of PARSEME guidelines for MVCs that the first verb (V-dep) should be non-finite. Therefore, passives clearly are not a case of VMWEs in Hindi.

Annotating MVCs was a little challenging as there is no direct relation in UD framework that can identify these verb+verb constructions. Further, we have to avoid constructions like modals and passives to be falsely tagged. Therefore, we have applied a number of rules to identify MVCs.

We have first filtered verbs that were tagged as 'VM' (main verb) for their xpos and are followed by auxiliary verbs (tagged as VAUX). Since, VAUX in all of these annotations includes any verb that has not been annotated as the main verb of the sentence, we decided to use a list of commonly used auxiliaries in Hindi including copulas, progressive marker, modals, and /vala/ to filter any false positive MVC cases, thereby also resolving the issue of modal constructions being tagged as MVCs. We have also filtered main verbs for any tense, aspect, and agreement inflections resulting in verbs that are in their base or infinitive form to avoid tagging of passives.

MVCs have also been added to the existing annotations according to the guidelines. If it already exists then we do not make any changes. It was followed by manual adjudication of the data to remove any false positive cases.

On comparing with original numbers (c.f Table 1), the total number of MVCs has dropped to 269. The reason is the removal of modals and passives from the data.

3.2.3. Stacked VMWEs

In Section 2.3 we have mentioned that we are not introducing any new label for Stacked VMWEs. As discussed, Stacked VMWEs shows recursive use of different types of multiword expressions occurring as a single predicate. Therefore, they can be easily retrieved using existing annotations for LVCs, MVCs, and causatives. For instance, as illustrated in Figure 1, we can extract by looking for verbs that are annotated for both LVCs and MVCs. Table 3 shows the frequency of stacked VMWEs. Also, note that since PARSEME has not reported the numbers for Stacked VMWEs in their previous editions of the language we have kept it as null.

Data	**LVC.full +MVC**	**LVC.cause +MVC**
PARSEME	null	null
New	61	1

Table 3: Number of Stacked VMWEs in the existing PARSEME corpus as compared to the New corpus.

The above table also highlights the fact that stacking of one VMWE onto another increases the complexity of the predicates and therefore occurs less frequently when compared to other VMWEs. As we can see that there was only one instance of LVC+causative+MVC kind of expression.

3.3. Verbal Idioms

Multiword Expressions are known for their non-compositionality with VIDs being the most diverse category such that detection of VIDs by automatic means was challenging. There were two types of issues. First, when a VID was tagged with a different VMWE category. Second, when an expression from another VMWE category was annotated as VID. Therefore, we have annotated them manually using PARSEME guidelines (Ramisch et al., 2020). These led to changes in the overall numbers of VIDs. As we can see in Table 4 the numbers have increased after the reannotation of the data especially after identifying the miscategorized VIDs.

4. Results and Conclusion

The main aim of this study was to enhance the existing PARSEME Hindi corpus by expanding its scope

Data	VID
PARSEME	61
New	74

Table 4: Number of VIDs in the existing PARSEME corpus as compared to the New corpus.

to other phenomena that results in the formation of different types of multiword expressions. Towards this goal we have proposed to annotate causatives via a morphological feature and to extract stacked VMWEs by using the existing annotations of other VMWEs. The new corpus now have the following categories – VID, LVC.full, LVC.cause, MVC, Causative, and Stacked VMWE.

Further, the results show that Hindi frequently employs VMWEs as shown in Figure 5. LVC.full are more common where as stacked VMWEs are rarer.

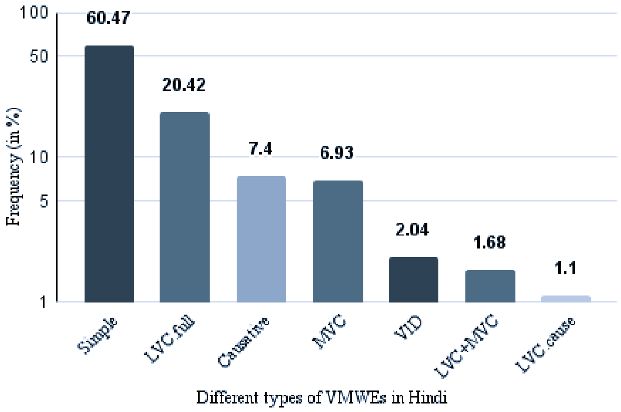


Figure 5: Frequency distribution of Hindi verbs in the new corpus.

Both Stacked VMWEs as well as causatives are infrequent as compared to other VMWE categories in all types of Hindi corpora. Our survey of corpora from other genres e.g., the Hindi TimeBank (Goel et al., 2020) and the IIT Delhi Dialogue Corpus for Hindi (Pareek et al., 2023) shows that Stacked VMWEs and causatives are consistently used (although they are relatively infrequent). We believe it is important to include these categories in the annotation framework to have a complete picture of VMWEs in Hindi.

Another goal of this study was to refine the existing annotations. For this, we have conducted a survey and identified a number of issues in the corpus. We have added annotations for the missing cases across different categories of VMWEs and removing any erroneous cases. The refinement process involved a combination of an automatic and manual annotation followed by adjudication. In case of automatic annotations we have described a method using UD framework to annotate some of the categories.

5. References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Rajesh Bhatt, Tina Bögel, Miriam Butt, Annette Hautli, and Sebastian Sulger. 2011. Urdu/hindi modals. *Proceedings of the LFG11 conference*, pages 47–67.

Miriam Butt. 1995. *The structure of complex predicates in Urdu*. Center for the Study of Language (CSLI).

Miriam Butt. 2010. The light verb jungle: Still hacking away. *Complex predicates in cross-linguistic perspective*, pages 48–78.

Miriam Butt and Tracy Holloway King. 2006. Restriction for morphological valency alternations : The Urdu causative. In Miriam Butt, editor, *Intelligent Linguistic Architectures : Variations on Themes by Ronald M. Kaplan*, number 179 in CSLI lecture notes, pages 235–258. CSLI Publications, Stanford, California.

Miriam Butt, Tracy Holloway King, and John T Maxwell III. 2003. Complex predicates via restriction. In *Proceedings of the LFG03 Conference*, pages 92–104.

Miriam Butt, Tracy Holloway King, and Gillian Ramchand. 2008. Complex predication: How did the child pinch the elephant. *Reality Exploration and Discovery: Pattern Interaction in Language & Life. A Festschrift for KP Mohanan*, pages 231–256.

Miriam Butt and Gillian Ramchand. 2005. Complex aspectual structure in Hindi/Urdu. In *The Syntax of Aspect*, pages 117–153. Oxford University Press Oxford.

Pranav Goel, Suhan Prabhu, Alok Debnath, Priyank Modi, and Manish Shrivastava. 2020. Hindi TimeBank: An ISO-TimeML annotated reference corpus. In *16th Joint ACL - ISO Workshop on Interoperable Semantic Annotation PROCEEDINGS*, pages 13–21, Marseille. European Language Resources Association.

Benu Pareek, Mudafia Zafar, Karan Yadav, Meghna Hooda, Ashwini Vaidya, and Samar Husain. 2023. The IIT Delhi Dialogue Corpus for Hindi. In Preparation.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Ashwini Vaidya, Sumeet Agarwal, and Martha Palmer. 2016. Linguistic features for Hindi light verb construction identification. In *International Conference on Computational Linguistics*.

Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities

Cvetana Krstev, Ranka Stanković, Aleksandra Marković, Teodora Mihajlov
Association for Language Resources and Technologies, Univ. of Belgrade, F. of Mining and Geology,
Institute for the Serbian Language SASA, Association for Language Resources and Technologies
cvetana@jerteh.rs, ranka@rgf.bg.ac.rs, malexa39@gmail.com, teodoramihajlov@gmail.com

Abstract

This paper presents the work in progress on ELEXIS-sr corpus, the Serbian addition to the ELEXIS multilingual annotated corpus (Martelli et al., 2023), comprising semantic annotations and word sense repositories. The ELEXIS corpus has parallel annotations in ten European languages, serving as a cross-lingual benchmark for evaluating low and medium-resourced European languages. The focus in this paper is on multiword expressions (MWEs) and named entities (NEs), their recognition in the ELEXIS-sr sentence set, and comparison with annotations in other languages. The first steps in building the Serbian sense inventory are discussed, and some results concerning MWEs and NEs are analysed. Once completed, the ELEXIS-sr corpus will be the first sense annotated corpus using the Serbian WordNet (SrpWN). Finally, ideas to represent MWE lexicon entries as Linguistic Linked-Open Data (LLOD) and connect them with occurrences in the corpus are presented.

Keywords: multiword expression, named entity, word sense disambiguation, sense repository, LLOD

1. Introduction

Even in the current era of neural language models, there is a high demand for high-quality, openly accessible corpora that are annotated with senses, especially for training and evaluating semantically related NLP tasks, like word sense disambiguation (WSD) and natural language understanding (NLU) (Pedersen et al., 2023b). Despite many efforts in the field over the past decades, such corpora are still scarce for many languages with limited resources, including Serbian. This scarcity is caused not only by the lack of freely available sense inventories, which are necessary for these tasks, but also by the complexity and cost of compiling annotations since it requires substantial manpower, preferably from experienced linguists or lexicographers. For Serbian, the availability of curated dictionaries for such use is limited, and not even subsets for particular corpora annotation are available.

The paper is structured as follows: In Section 2 we give an account of related work, and continue by presenting in Section 3 the ELEXIS-WSD dataset, its extension with Serbian data and its basic annotation layers prior to the semantic annotation. Section 4 discusses the annotation of MWEs and NEs in the ELEXIS-WSD as well as in its Serbian extension. The building of the sense inventory for Serbian and the role of MWEs and NEs in it are presented in Section 5. Possible ideas for publishing dictionaries of MWEs as LLOD and associating its entries with corresponding occurrences in the corpus are developed in Section 6. Finally, in Section 7 we conclude and discuss open questions, potential future research, and development.

2. Related work

A semantic concordance is a textual corpus and a lexicon, combined so that every substantive word in the text is linked to its appropriate sense in the lexicon (Miller et al., 1993). The popularity of SemCor (Landes et al., 1998), one of the initial sense-annotated English corpora based on the Princeton WordNet sense inventory (Fellbaum, 1998) inspired the NLP community to build sense-annotated corpora for many languages. Exploiting parallel texts in the creation of multilingual semantically annotated resources produced the MultiSemCor Corpus (Bentivogli and Pianta, 2005). Another important multilingual sense annotated corpus is the Ontonotes (Weischedel et al., 2011), that uses the WordNet for sense annotations of the English part, whereas the Chinese and Arab parts base the sense annotations on various lexical sources.

The semiautomatic approaches to sense annotation were applied to overcome the scarcity of such data sets. The OneSec are sense-annotated corpora for word sense disambiguation in multiple languages and domains (Scarlini et al., 2020) that consist of Wikipedia texts containing between 1.2 and 8.8 M sense annotations of nouns per language.

The FrameNet project (Baker et al., 1998), based on the idea of describing lexical items through semantic frames, produces semantic frames (which contain information about the semantic and syntactic valence of words). Target words are mostly nouns, adjectives, and verbs. Every frame and frame element is accompanied by a set of representative sets of manually annotated corpus attestations, and for every frame, the set of relations it enters is presented. Lexical databases based on the FrameNet principles were (or are being) built for several languages. The Salsa project (Burchardt et al., 2009) produced a German lexicon based on the FrameNet semantic frames and annotated a large German newswire corpus.

The English part of Ontonotes was annotated with verbal MWEs (Kato et al., 2018). The main outcome of the COST action PARSEME were unified annotation guidelines, and a corpus of over 5.4 million words and 62 thousand annotated VMWEs in 18 languages (Savary et al., 2018). Development was continued afterward with the inclusion of more languages and the enlargement of corpora for existing languages. The current edition of PARSEME corpus[1] contains 26 languages, including Serbian (Savary et al., 2023). The expansion of MWE annotations to nominal and other MWEs is the task within COST action UNIDIVE[2].

Named Entity Recognition (NER) enables the identification and classification of key information in text. The most frequently annotated classes are persons, locations, and organizations, but for a deeper text understanding identification of events, roles, time, measures, etc. is also necessary. In addition to that, named entity linking (NEL), also known as disambiguation, normalization, or entity resolution, involves aligning a textual mention of a named entity to an appropriate entry in a knowledge base, assigning a unique identity to mentioned entities.

3. The extension of ELEXIS-WSD

ELEXIS-WSD is a parallel sense-annotated corpus in which content words (nouns, adjectives, verbs, and adverbs) have been assigned senses for 10 languages: Bulgarian (BG), Danish (DA), English (EN), Spanish (ES), Estonian (ET), Hungarian (HU), Italian (IT), Dutch (NL), Portuguese (PT), and Slovenian (SL).[3] The list of sense inventories is based on WordNet for DA (Pedersen et al., 2023a), EN, IT, NL, Wiktionary is used for ES, and national digital dictionaries are used for BG, ET, HU, PT, and SL (Federico et al., 2021).

In order to join this task and obtain the Serbian corpus as a part of the future edition of the sense repository being developed within WG2.T2 of the UniDive, the set of sentences from WikiMatrix[4] in EN was translated automatically (Google translation) into SR. We opted for the automatic translation in order to fasten the process, with the full awareness of the need to manually check the translation afterwards. This process was highly demanding, in terms of time and manpower, but it was an unavoidable step for getting high-quality dataset. A few (eight precisely) Serbian native speakers checked the Serbian sentence set thoroughly, in order to avoid literal or incorrect translation, and after that sentences were read carefully once again to resolve different issues: literally or incorrectly translated MWEs, unresolved references in the text (pronouns, e.g., in SR differ for gender, number, and case, and if the pronoun refers to an NP from the previous context, its reference had to be checked to choose the right morphological form); besides, it was necessary to check phonetic transcriptions of names (particularly personal ones), since in SR proper names are not written in the original form (the second reading and issue-resolving was done by two people). The process was time-consuming because of the very nature of the set – sentences are out of context, full of terms from different scientific areas, many of which are MWEs), their content is of encyclopedic sort, and often it was necessary to read the original document in English and/or some other language to understand the meaning and represent it correctly in SR.

After this process, the set was automatically tokenized, lemmatized, and POS-tagged (Stanković et al., 2020; Stanković et al., 2022). The outcomes of all these automatic procedures are being manually corrected. Results show that 2024 sentences in Serbian dataset have 25,478 word forms, content words tagged as: NOUN – 7,198 (diff. 2,413), PROPN – 1,552 (diff. 1,057), ADJ – 3,291 (diff. 1,256), VERB – 3,121 (diff. 913), ADV – 900 (diff. 287).[5] Tasks that remain to be done include the annotation of MWEs and NEs (the first results are presented in the following section), the syntactic annotation, and linking with the sense repository (the first results are presented in Section 5).

4. MWEs and NEs in WSD

In this section, we are focusing on the annotation of MWEs and NEs in the ELEXIS-WSD and in its Serbian extension. As it will be shown, the number of MWEs and NEs annotated in 10 language sentence sets was not even, probably due to different resources used for their annotation.

[1] https://gitlab.com/parseme/

[2] https://unidive.lisn.upsaclay.fr/

[3] https://www.clarin.si/repository/xmlui/handle/11356/1842

[4] https://ai.meta.com/blog/wikimatrix/

[5] This figures are not final since the annotation is presently being double-checked and harmonized with UD.

In order to compare MWEs and NEs occurring in the whole repository, MWEs and NEs in each of 10 languages were automatically translated into SR (as phrases, not word-to-word), and the number of the same translations obtained by translating MWEs from different languages was calculated. MWEs/NEs were automatically translated into Serbian in order to facilitate the comparison with MWEs/NEs retrieved in the Serbian set. Note that the translation of a MWE into SR need not be a MWE. For instance, *prime minister* (EN) → *premijer* (SR).

4.1. MWEs in ELEXIS-WSD

The number of annotated MWEs in the initial WSD repository is presented in Figure 1). The blue columns present the number of unique lemmas in the WSD, while the orange columns present the number of unique senses. This graphic shows that the numbers of MWEs in the WSD repository differ significantly between languages.

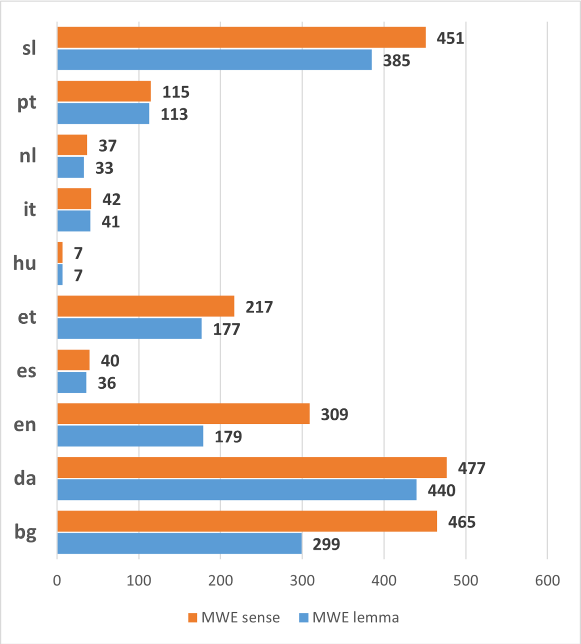


Figure 1: Number of MWEs in the repository – a total of 1,710 MWEs in 10 languages

.

Figure 2 shows that 1,412 different translations were obtained by translating a total of 1,710 MWEs. One international MWE appeared in 6 language sets, *lingua franca*. One of 14 MWEs translated from 4 languages into one SR term was *očekivano trajanje života* (SR): *life expectancy* (EN), *expectativa de vida* (PT), *pričakovana življenjska doba* (SL), *oodatav eluiga* (ET).[6]

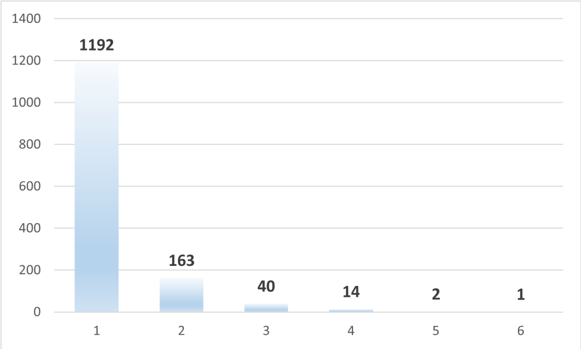


Figure 2: MWEs translations into Serbian obtained by translating from 10 languages – no translation was obtained by translating from more than 6 languages

.

4.2. NEs in ELEXIS-WSD

The number of annotated NEs, without information about specific NE types, is presented in Figure 3 (blue columns present the number of unique lemmas, while orange columns present the number of unique senses).

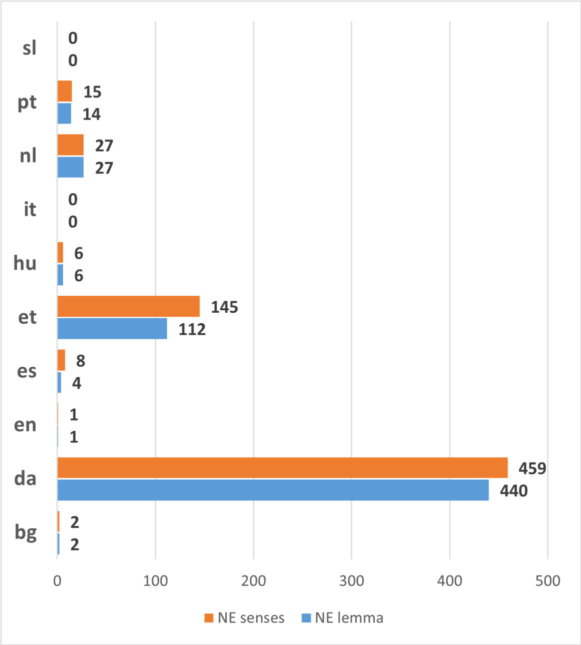


Figure 3: Number of NEs in the repository – a total of 606 NEs in 10 languages

.

Named entities were not systematically annotated in all language datasets (for example, (SL) and (IT) sets have no NE annotated at all, while

[6] The automatic translation was not literal, as demonstrated by the example *srednja škola* (SR) 'lit. middle school' ↔ *high school* (EN); on the other hand it was not always accurate, as demonstrated by *srednja škola* (SR) ↔ *visoka šola* (SL).

some languages have many of them), resulting in 526 translations from a total of 606 NEs. The most frequent NE was *Grčka*, translated from four languages: *Grækenland* (DA), *Grecia* (ES), *Kreeka* (ET), *Grécia* (PT), followed by NEs translated from three languages, one of which is *SAD*: *USA* (EN), *ZDA* (SL), *EUA* (PT). Figure 4 presents the number of translations into Serbian.

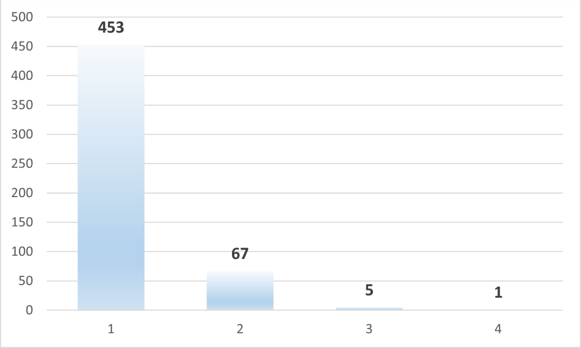


Figure 4: NEs translations into Serbian obtained by translating from 10 languages – no translation was obtained by translating from more than 4 languages

.

4.3. Annotation of MWEs and NEs in the Serbian dataset

The pipeline for preparation and annotation of the Serbian set of 2,024 sentences is presented in Figure 5 – green color boxes and the closed locker symbol represent the finished tasks, pink color boxes and the open locker symbol designate the work in progress, mostly in the evaluation phase, while the pending tasks or tasks in their initial phase are represented by lilac boxes.

The Serbian set of 2,024 sentences was automatically annotated using four different resources and tools:

- The e-dictionary of non-verbal MWEs was used for the annotation of such MWEs. This dictionary was built on the same principles used for building the e-dictionary of simple words for Serbian. The inclusion of MWEs in this dictionary was based on several rather loose criteria: their appearance in some general, terminological or phraseological dictionary of Serbian as well as SrpWN, the frequency of their occurrence in corpora of Serbian, and the intuition of the resource author. The application of this resource to the Serbian sentence set resulted in 529 annotations (339 different) (Krstev et al., 2013). Among them were 351 (249) nominal MWEs, 133 (70) proper nouns, 44 (19) adverbial, and one adjectival.

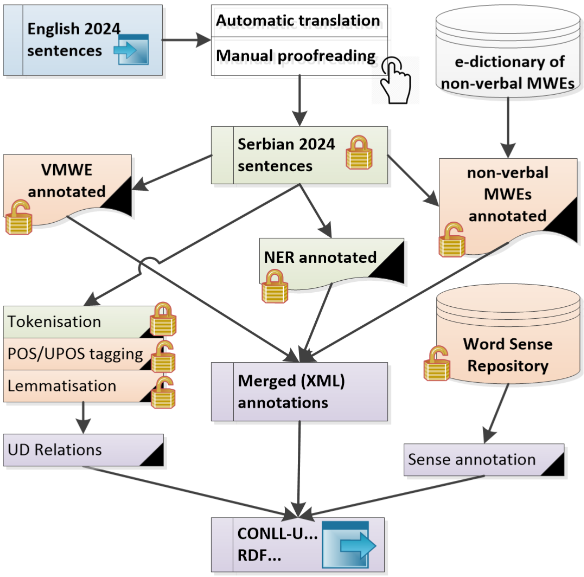


Figure 5: The pipeline for the preparation of the Serbian dataset

.

- A system for the Named Entity Recognition (NER) based on e-dictionaries and rules annotated 2,006 occurrences of NEs (Krstev et al., 2014). Numbers of recognized NEs per class are presented in Table 1. Some multiword named entities, particularly organization (ORG) and geopolitical names (TOP), are recognized both by dictionaries and the NER system.

- A system for the recognition of verbal MWEs based on e-dictionaries, rules, and the repertoire of VMWEs annotated in the Serbian part of the PARSEME Corpus Release 1.3 (Savary et al., 2023) annotated 230 occurrences of VMWEs (98 different), distribution by type: IRV – 174 (62), LVC.full – 35 (21), VID – 13 (10), and LVC.cause – 8 (5).

- A system for the recognition of adjectival and verbal similes is based on a set of more than 600 adjectival and more than 300 verbal similes. It can retrieve different variances of these similes, both in lexica and structure (Krstev et al., 2023). The previous research established that in literary texts an average of 2.2 adjectival similes can be expected per 10,000 words of a text; however, this system in the Serbian sentence set did not retrieve even a single one (Krstev, 2021).

The accuracy of MWE/NE recognition, recall, and precision will be determined during the next step, when senses will be associated with simple- and multi-word units. Previous evaluations of used systems for the recognition of NEs and MWEs (Krstev et al., 2013; Šandrih et al., 2019) have shown that these systems prioritize precision over recall, which means that in the later stages of processing, through comparison with annotations in datasets for other languages and manual evaluation, new entities will be annotated. It is to be expected that the assignment of senses will reveal some additional MWEs and NEs. This, in turn, will enable the enhancement of used resources and procedures.

Tag	№	Tag	№
PERS	329	TIME	372
TOP	448	AMOUNT	169
ORG	126	MEASURE	62
DEMONYM	244	PERCENT	51
ROLE	175	MONEY	12
EVENT	18	**Total**	2,006

Table 1: Recognized NEs by classes.

4.4. The comparison of MWEs and NEs across languages

Our initial comparison of MWEs and NEs annotated in the WSD repository and in the Serbian sentence set (ELEXIS-SR) was based on their automatic translation to SR, as explained in Subsection 4. This was not ideal, since in several cases the translation was not appropriate: e.g., the SR highly polysemous verb *dovesti* was obtained as a translation equivalent of two VMWEs from two languages, appearing in two unrelated sentences: *dado lugar, 'leed to'* (ES: 700) and *tog med,'take in'* (DA: 148). Once the automatic translation was checked, as actual equivalents of these VMWEs in ELEXIS-SR appeared to be *primiti* (148) ('take in' in ELEXIS-EN), and *dovesti* (700) ('leed to' in ELEXIS-EN), the translated verb itself.

On the other hand, in many cases automatic matches were good: e.g., the SR translation *bruto domaći proizvod* 'gross domestic product' was obtained from MWEs in four languages: *gross domestic product* (EN), *produto interno bruto* (PT), *bruto domači proizvod* (SL), *sisemajandus koguprodukt* (ET), all occurring in the same sentence – 1258. In the corresponding sentence in ELEXIS-SR the translated term *bruto domaći proizvod* was used and annotated as MWE. In all mentioned languages these terms were also annotated as MWEs (in ELEXIS-SL only its part is annotated: *domači proizvod*).

In other cases, the translation was good, it was used in ELEXIS-SR, but it was not annotated in it because it was missing in the used resources. This was the case for *prirodna selekcija*, translated from *natural selection* (EN), *seleção natural* (PT), *naravni izbor* (SL), used in sentence 1560 in ELEXIS-SR, but not annotated in it. This case of missing annotations occurs in other languages as well. E.g., equivalents for *gross domestic product* are MWEs in BG, ES, HU, NL, but yet are not annotated. In IT an acronym was used instead, and in DA a compound.

Having all this in mind, the overall results of the comparison are as follows: out of 653 non-verbal MWEs occurrences (384 lemmas) annotated in ELEXIS-SR, 116 MWE lemmas occurred in at least one language set in WSD; out of 228 VMWE occurrences (99 lemmas) annotated in ELEXIS-SR, 11 lemmas occurred in at least one language set; only 93 NEs annotated in ELEXIS-SR were annotated as MWE or PROPN in WSD (maybe due to the poor lemmatization, automatic translation and linking of proper names).

5. Sense repository

Since there is no freely available digital descriptive dictionary of the Serbian language, the Serbian sense repository will be based on the Serbian WordNet SrpWN (Stanković et al., 2018). ELEXIS sense repository for English, which is also based on the Princeton WordNet (PWN), has 16,106 entries, each assigned with its internal identifier. Since the WordNet interlingual index is not available in the ELEXIS-EN sense repository, we aligned PWN synsets with the ELEXIS-EN sense repository entries by comparing their definitions. This process yielded 13,703 matches.

Subsequently, synsets from this subset were aligned with the Serbian WN containing 25,322 synsets, which revealed that there were 5,997 matches. Finally, the subset missing from the list of 13,703 synsets was compared with sentence annotations in ELEXIS-EN, which revealed that the "urgent" first step is to fill the gap with 2,130 synsets. After the automatic translation of this list of synonyms and their definitions from the PWN using Google API and OpenAI services, the obtained list of Serbian candidates was expanded using several other lexical resources compiled in previous research. Postediting the list of synonym set candidates and their definitions is an ongoing activity.

The further analysis showed that from 437 MWEs annotated in ELEXIS-SR (see Subsection 4) – 339 non-verbal and 98 verbal – 171 (39%) were found in the Serbian WordNet. Moreover, some of them occur in 2 or more synsets. Table 2 gives the total number of senses and the number of lemmas (literals) per MWE type. For instance, *komunikacioni sistem* 'communication system' can refer to a '(def.) system for communicating' or to

Group	Type	Senses	Lemma
MWE	NOUN	100	94
MWE	PNOUN	35	32
VMWE	IRV	80	42
VMWE	LVCfull	1	1
VMWE	VID	2	2

Table 2: Annotated MWEs in ELEXIS-SR retrieved in SrpWN per type.

'(def.) a facility consisting of the physical plants and equipment for disseminating information.' The VMWE with the highest number of different meanings is the reflexive verb *pojaviti se*: 'to appear' (to come in sight), 'to come up' (of celestial bodies), 'to come out' (be issued), 'to originate' (come into existence), 'to arise' (result or issue). A proper noun having two senses is *Novi Zeland* 'New Zealand', referring to a country and an island (same as in the PWN). Besides reflexive verbs, one light verb construction is recorded in the subset of SrpWN potentially interesting for our research – *dati ostavku* 'give resignation' – and two verbal idioms – *uzeti u obzir* 'take into consideration' and *voditi računa* 'take care'.

There are 533 synsets from the SrpWN which contain MWE literals that correspond to the synsets in the PWN used to annotate senses in ELEXIS-EN. Among them are 476 that were not annotated by our tools. Naturally, a number of them does not appear in ELEXIS-SR. For instance, *vodonik* 'hydrogen' appears in sentence 272, but its synonym *atomski broj 1* 'atomic number 1' does not. In some cases the Serbian correct translation avoids the use of a certain expression used in ELEXIS-EN, e.g. the sentence 597 ends with '...the best time being this summer.' while the same sentence in Serbian ends with *...najbolje ovog leta.* 'best this summer', avoiding the use of Serbian counter terms for 'time' in the sense 'a suitable moment', MWEs *pravi trenutak* or *pravi čas*. Finally, there are MWEs, like *merna jedinica* 'measurment unit', used in the sentence 735, which were not annotated since they were not yet recorded in lexical resources used for the annotation.

6. Linking MWEs and corpora

A holistic presentation of MWEs in lexicons and linking their entries with occurrences in a corpus is still an open question. We are considering the use of LLOD for interlinking MWE lexicon entries with their occurrences in corpora. Two options will be taken into account: Ontolex-lemon[7] (with Lexicog module) and DMLex[8]. Ontolex-lemon is widely used community standard for machine-readable lexical resources in the context of RDF, Linked Data, and Semantic Web technologies (McCrae et al., 2017). DMLex is a standard for structuring (human-oriented) dictionaries, which is published by LEXIDMA, a technical committee under OASIS, an organisation which oversees the production of open standards in the IT industry.

The following example gives an idea of how the lexical entry for MWE *fast food* can be represented in RDF along with its translations – in this case in Serbian *brza hrana* – and how links between senses and Wikidata entries are realized.

```
:le_fast_food
  a ontolex:LexicalEntry,
    ontolex:MultiwordExpression;
  ontolex:canonicalForm
    [ontolex:writtenRep
    "fast food"@en];
  lexinfo:partOfSpeech lexinfo:noun;
  ontolex:sense
    [ontolex:reference
<https://www.wikidata.org/wiki/Q81799>];
  decomp:constituent :cm_fast;
  decomp:constituent :cm_food;
  rdf:_1 :le_fast; # lexical
  rdf:_2 :le_food. # entries

# component of cannonical form
:cm_food  a decomp:Component;
  decomp:correspondsTo :le_food.
  …
:le_brza_hrana a ontolex:LexicalEntry,
    ontolex:MultiwordExpression;
  ontolex:canonicalForm
    [ontolex:writtenRep
    "brza hrana"@sr];
  …

# simplified naming
:tranSetEN-SR vartrans:trans
 fast_food-ensns-brza_hrana-srsns .
:fast_food-ensns
 a ontolex:LexicalSense ;
 ontolex:isSenseOf :le_fast_food .
:brza_hrana-srsns
 a ontolex:LexicalSense ;
 ontolex:isSenseOf :le_brza_hrana .
:fast_food-ensns-brza_hrana-sns-trans
 a vartrans:Translation ;
 vartrans:source :fast_food-ensns ;
 vartrans:target :brza_hrana-srsns .
```

The OntoLex-FrAC[9] vocabulary implements the lexicon-corpus interface (Barbu Mititelu et al.,

[7] https://www.w3.org/2016/05/ontolex

[8] https://www.lexiconista.com/dmlex/;https://docs.oasis-open.org/lexidma/dmlex/v1.0/csd02/dmlex-v1.0-csd02.html

[9] The current draft version of the FrAC specification is found under https://github.com/ontolex/frequency-attestation-corpus-information/

2024) with corpus information to support corpus-driven lexicography and the inclusion of corpus evidence (attestations). A sentence number 823 in English: “It can be made at home or bought from fast food shops.” and in Serbian “Može se napraviti kod kuće ili kupiti u prodavnicama brze hrane.” illustrates this in the following example:

```
:le_fast_food
 frac:attestation [
 frac:quotation "It can be made at
 home or bought from fast food
 shops."@en;
 frac:observedIn :EWSD].

:le_brza_hrana
 frac:attestation [
 frac:quotation "Može se napraviti
 kod kuće ili kupiti u prodavnicama
 brze hrane."@sr;
 frac:observedIn :EWSD-ext].
```

The cross-lingual analysis of idiosyncratic constructions can be supported by publishing aligned and annotated corpus data as Linked Data employing community standards such as the NLP Interchange Format (NIF) (Hellmann et al., 2012) and CoNLL-RDF (Chiarcos and Fäth, 2017; Chiarcos and Glaser, 2020), a minimal NIF subset designed for compatibility with tab-separated formats used in NLP (“CoNLL”), Universal Dependencies (“CoNLL-U”) and Parseme (“Parseme-TSV”). The sense repository should be probably published using Ontolex-lemon. The first ideas about leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora are given in (Stanković et al., 2023).

7. Future Work

The development of the Serbian sentence set is a work in progress, as represented in Figure 5: translation and tokenization are done, POS tagging and lemmatization checking are in the final phase, and word sense inventory is being prepared. The syntactic annotation is still pending as well as the development of a LLOD dictionary.

Our future research will give special attention to the annotation of MWEs and NEs in the ELEXIS-SR. On the one hand, we will coordinate our work with the other research groups, primarily groups dealing with ELEXIS, Parseme, UD and UniDive activities. On the other hand, having in mind that the meticulously prepared set ELEXIS-SR will be used for various purposes, we plan to publish its various editions, e.g. annotating NEs using a large set of classes and sub-classes. One important research path will be the production of precise guidelines for distinguishing MWEs from NEs, as well as explicating differences in the notion of a MWE in the Serbian e-dictionaries, Parseme/UniDive and WordNet (e.g. MWEs *pravi trenutak* or *pravi čas* ‘time (a suitable moment)’ would probably not be considered a nominal MWE for Parseme/UniDive,[10] that is, they would not pass the prescribed sequence of tests).

The future research goal is the comparative analyses of MWEs and NEs in ELEXIS multilingual set both from the linguistic and NLP point of view.

Acknowledgements

This research was supported by the Ministry of Science, the Republic of Serbia, #GRANT 451-03-65/2024-03/ 200126,#GRANT 451-03-66/2024-03/200174 and COST ACTION CA21167 - Universality, Diversity, and Idiosyncrasy in Language Technology (UniDive).

8. Bibliographical References

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet Project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Verginica Barbu Mititelu, Voula Giouli, Stella Markantonatou, Ivelina Stoyanova, Petya Osenova, Kilian Evang, Daniel Zeman, Simon Krek, Carole Tiberius, Christian Chiarcos, and Ranka Stankovic. 2024. Multiword expressions between the corpus and the lexicon: Universality, idiosyncrasy and the lexicon-corpus interface. In *Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)*.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(3):247–261.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2009. FrameNet for the semantic analysis of German: Annotation, representation and automation. *Multilingual FrameNets in Computational Lexicography: methods and applications*, 200:209–244.

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *Language, Data, and Knowledge: First International Conference, LDK 2017,*

[10] The guidelines for non-verbal MWEs are still in preparation.

Galway, Ireland, June 19-20, 2017, Proceedings 1, pages 74–88. Springer.

Christian Chiarcos and Luis Glaser. 2020. A tree extension for CoNLL-RDF. In *Proceedings of the 12th LREC*, pages 7161–7169.

Martelli Federico, Navigli Roberto, Krek Simon, Kallas Jelena, Gantar Polona, Veronika Lipp, Tamás Váradi, András Győrffy, and László Simon. 2021. Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In *Proceedings of the eLex 2021 conference*, pages 377–395. Lexical Computing.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Marcus Nitzschke. 2012. NIF Combinator: Combining NLP Tool Output. In *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, 2012. Proceedings 18*, pages 446–449. Springer.

Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Construction of large-scale English verbal multiword expression annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Cvetana Krstev. 2021. White as snow, black as night – similes in old serbian literary texts. *Infotheca - Journal for Digital Humanities*, 21(2):119–135.

Cvetana Krstev, Ivan Obradović, Ranka Stanković, and Duško Vitas. 2013. An approach to efficient processing of multi-word units. *Computational Linguistics: Applications*, pages 109–129.

Cvetana Krstev, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, 24(2):473–489.

Cvetana Krstev, Ranka Stanković, and Aleksandra Marković. 2023. Multiword expressions – comparative analysis based on aligned corpora. In *Book of Abstracts of the UniDive 1st general meeting, 16-17 March 2023, Paris-Saclay University, France*.

Shari Landes, Claudia Leacock, and Christiane Fellbaum. 1998. Building semantic concordances. In Christian Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 199–216. MIT Press.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: Development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Bolette Pedersen, Sanni Nimb, Sussi Olsen, Thomas Troelsgård, Ida Flörke, Jonas Jensen, and Henrik Lorentzen. 2023a. The DA-ELEXIS corpus - a sense-annotated corpus for Danish with parallel annotations for nine European languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 11–18, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Bolette Sandford Pedersen, Sanni Nimb, Sussi Olsen, Thomas Troelsgård, Ida Flörke, Jonas Jensen, and Henrik Lorentzen. 2023b. The DA-ELEXIS Corpus – a Sense-Annotated Corpus for Danish with Parallel Annotations for Nine European Languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 11–18.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, and et al. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten van Gompel, et al. 2018. PARSEME multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5905–5911.

Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020.

Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for Serbian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3954–3962, Marseille, France. European Language Resources Association.

Ranka Stanković, Christian Chiarcos, and Milica Ikonić Nešić. 2023. Leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora. In *Book of Abstracts of the UniDive 1st general meeting, 16-17 March 2023, Paris-Saclay University, France*.

Ranka Stanković, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. 2018. Resource-based WordNet Augmentation and Enrichment. In *Proceedings of the Third International Conference Computational Linguistics in Bulgaria (CLIB 2018)*, pages 104–114, Sofia, Bulgaria. Institute for Bulgarian Language “Prof. Lyubomir Andreychin”, Bulgarian Academy of Sciences.

Ranka Stanković, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. Parallel Bidirectionally Pretrained Taggers as Feature Generators. *Applied Sciences*, 12(10).

Ralph Weischedel, Eduard Hovy, Marcus Mitchell, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural LanguageProcessing and Machine Translation: DARPA GlobalAutonomous Language Exploitation*.

Branislava Šandrih, Cvetana Krstev, and Ranka Stanković. 2019. Development and evaluation of three named entity recognition systems for serbian - the case of personal names. In *Natural Language Processing in a Deep Learning World – Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1061–1068. INCOMA Ltd.

9. Language Resource References

Martelli, Federico and Navigli, Roberto and Krek, Simon and Kallas, Jelena and Gantar, Polona and Koeva, Svetla and Nimb, Sanni and Sandford Pedersen, Bolette and Olsen, Sussi and Langemets, Margit and Koppel, Kristina and Üksik, Tiiu and Dobrovoljc, Kaja and Ureña-Ruiz, Rafael and Sancho-Sánchez, José-Luis and Lipp, Veronika and Váradi, Tamás and Győrffy, András and Simon, László and Quochi, Valeria and Monachini, Monica and Frontini, Francesca and Tiberius, Carole and Tempelaars, Rob and Costa, Rute and Salgado, Ana and Čibej, Jaka and Munda, Tina and Kosem, Iztok and Roblek, Rebeka and Kamenšek, Urška and Zaranšek, Petra and Zgaga, Karolina and Ponikvar, Primož and Terčon, Luka and Jensen, Jonas and Flörke, Ida and Lorentzen, Henrik and Troelsgård, Thomas and Blagoeva, Diana and Hristov, Dimitar and Kolkovska, Sia. 2023. *Parallel sense-annotated corpus ELEXIS-WSD 1.1*. The Jožef Stefan Institute. Slovenian language resource repository CLARIN.SI.

To Leave No Stone Unturned: Annotating Verbal Idioms in the Parallel Meaning Bank

Rafael Ehren, Kilian Evang, Laura Kallmeyer
Heinrich Heine University Düsseldorf
Universitätsstr. 1, 40225 Düsseldorf, Germany
{rafael.ehren, kilian.evang, laura.kallmeyer}@hhu.de

Abstract

Idioms present many challenges to semantic annotation in a lexicalized framework, which leads to them being underrepresented or inadequately annotated in sembanks. In this work, we address this problem with respect to verbal idioms in the Parallel Meaning Bank (PMB), specifically in its German part, where only some idiomatic expressions have been annotated correctly. We first select candidate idiomatic expressions, then determine their idiomaticity status and whether they are decomposable or not, and then we annotate their semantics using WordNet senses and VerbNet semantic roles. Overall, inter-annotator agreement is very encouraging. A difficulty, however, is to choose the correct word sense. This is not surprising, given that English synsets are many and there is often no unique mapping from German idioms and words to them. Besides this, there are many subtle differences and interesting challenging cases. We discuss some of them in this paper.

Keywords: verbal idioms, semantic annotation

1. Introduction

Despite being one of the most discussed multiword expression (MWE) types, verbal idioms (VIDs) are surprisingly challenging to define. Actually, it seems to be easier to define them in terms of what they are not, as it is done by the PARSEME annotation guidelines (Ramisch et al., 2020)[1]. According to these guidelines, VIDs consist of a head verb and at least one lexicalized dependent which is neither a reflexive pronoun nor a particle. If the dependent is a verb or a noun, fine-grained tests need to be applied to discriminate the expression from multiverb expressions or light-verb constructions (LVCs). Another defining – and probably the most challenging – characteristic of an idiom is its non-compositionality, i.e. the meanings of its parts do not combine to form the meaning of the whole expression. However, since Nunberg et al. (1994), it is commonly acknowledged that there exists another dimension w.r.t. non-compositionality. We now make the distinction between decomposable and non-decomposable idioms. Both types are non-compositional, but for the former we can establish a mapping from its parts to their respective idiomatic meanings which in turn combine to form the meaning of the whole. Or, if we reverse the direction: We can decompose the idiomatic meaning and map these individual meanings to the components of the expressions.[2] This, however, is not possible for non-decomposable idioms whose meanings do not allow for this kind of distribution over their parts. For illustration, consider the following two classic examples:

(1) After a long interrogation the spy **spilled the beans.**

(2) After a long illness, he finally **kicked the bucket.**

Example (1) shows an instance of the idiom *spill the beans* which means ‘to reveal a secret’. We consider this decomposable because the individual meanings can be mapped to the different components of the expression: ‘reveal’ to *spill* and ‘secret’ to *beans*. Such a mapping does not exist for ‘kick the bucket’ in example (2) because the idiomatic meaning ‘to die’ cannot be decomposed into individual meanings.

Because of this behavior, non-decomposable idioms are more challenging when it comes to semantic annotation (and consequently semantic parsing) than decomposable ones. For the latter, there exists a one-to-one mapping from words to concepts, but not for the former. This might be the reason why they are often ignored during semantic annotation and receive a literal treatment. Consider the following example from the English partition of the Parallel Meaning Bank (PMB):

(3)

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=050_Cross-lingual_tests/030_Verbal_idioms__LB_VID_RB_

[2] Nunberg et al. (1994) spoke of idiomatically combining expressions, which reflects the initial direction of the analysis (starting from its parts), but since then the terminology changed in order to favor the other direction (starting from the whole expression).

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 115–124. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

x_1, e_1
pull.v.01(e_1), Agent(e_1, hearer), Theme(e_1, x_1), leg.n.01(x_1), Of(x_1, speaker)

Discourse representation structure (DRS) for English PMB sentence 01/1871 *Are you pulling my leg?* (not gold).

The non-decomposable idiom *pull sb's leg* has the meaning 'to tease sb', but in the DRS above it is treated literally as *leg* is a discourse referent (x_1) which it should not be. Thus, the DRS actually represents a leg pulling event which is not the desired analysis in this case.

The goal of this work is to improve the coverage of VIDs in the PMB, so that ultimately semantic parsers trained on its data can benefit from it. Furthermore, as a byproduct, we created a dataset of potentially idiomatic expressions (PIEs; Haagsma et al., 2020), since we also labeled instances of literal counterparts of VIDs. This will be further elaborated at the end of section 4.

The structure of the paper is as follows: First, we will discuss related work and the PMB. Then, we will detail the extraction of candidate sentences and the annotation process. Finally, we will present the results and discuss especially challenging cases before we draw our conclusions.

2. Related Work

Arguably the most well-known MWE corpora are the four editions (1.0–1.3) of the PARSEME corpus (Savary et al., 2015); (Ramisch et al., 2018, 2020; Savary et al., 2023). What sets them apart from other corpora is their scope and homogeneity: The PARSEME corpora consist of a large number of datasets from different languages that were all annotated for verbal MWEs according to the same annotation guidelines. PARSEME corpora are not sense annotated, but these guidelines are highly relevant to us, too, as we used their definitions of the different verbal MWE types to decide which candidate expressions to annotate.

A corpus that contains semantic annotation of MWEs is the STREUSLE corpus (Schneider and Smith, 2015). It is a 55,000 words English web corpus consisting of reviews which were annotated for MWEs, but without restrictions to specific kinds of syntactic constructions. Furthermore, it distinguishes between *strong* and *weak* expressions, the former being opaque idioms (*shoot the breeze*) while the latter are more transparent collocations (*traffic light*). On top of that, they added a level of supersenses which are the top-level hypernyms in the WordNet taxonomy. There is no explicit mention of decomposable and non-decomposable idioms, but the aformentioned *strong* expressions receive a supersense as a unit while *weak* ones do not. So it is probable that non-decomposable expressions received the appropriate treatment w.r.t. to supersense tagging. However, since there were no guidelines to differentiate decomposable and non-decomposable idioms, it is not unlikely that some of the former were annotated as strong and thus erroneously received a holistic treatment.

Sembanks (corpora with deep meaning representations) treat idioms in different ways. Abstract Meaning Representations (AMR; Banarescu et al., 2013) and Uniform Meaning Representations (UMR; van Gysel et al., 2021) are not lexically anchored, so usually introduce a single concept node for an idiom consisting of several words (Bonn et al., 2023). On the other hand, sembanks with lexical anchoring need explicit mechanisms for dealing with cases where the word-concept mapping is not one-to-one, such as idioms. For HPSG, such mechanisms have been proposed, e.g. by Richter and Sailer (2014), but not, to our knowledge, applied in sembanks such as LinGO Redwoods (Oepen et al., 2002).

3. The PMB

The Parallel Meaning Bank (PMB; Abzianidze et al., 2017, 2020) is a partially parallel corpus of text in English, German, Italian, and Dutch, with semantic annotations. These include WordNet senses (Fellbaum, 1998) and VerbNet semantic roles (Kipper Schuler, 2005), among others. All semantic annotation layers are integrated into a meaning representation language based on Discourse Representation Theory (Kamp and Reyle, 1993) which places more emphasis than other frameworks such as AMR on precisely representing the scope of quantifiers as well as modal and logical operators. The semantic representations in this formalism are called Discourse Representation Structures (DRS).

The PMB is built using a dynamic annotation methodology (Oepen et al., 2002) based on a strongly lexicalized theory of the syntax-semantics interface. Statistical models produce an initial syntactic analysis of each sentence using Combinatory Categorial Grammar (CCG; Steedman, 2001) as well as an assignment of semantic tags, roles, senses, etc. to tokens. These annotation layers are corrected by human annotators by adding constraints called *bits of wisdom*. Bits of wisdom are stored in a database so they can be automatically reapplied to the output of the new versions of the statistical models in the future. The result is then fed into a rule-based component named Boxer which assigns a partial meaning representation (λ-DRS) to each token and then computes a DRS for the entire sentence. Automatically preannotated documents are said to have 'bronze' status, documents with at least one bit of wisdom are 'silver', and documents marked as completely corrected by a human are 'gold'.

While the syntax-based annotation methodology of the PMB helps ensure consistency, it is challenged by multiword expressions where the mapping between lexical meanings and tokens is not one-to-one. Some types of verbal multiword expressions are already handled adequately. For example, in the verb-particle construction (4) and in inherently reflexive verbs (5), the meaning is assigned to the head, and the other element is treated as semantically empty. Decomposable verbal idioms as in (6) are treated by assigning each component a suitable non-literal meaning. Of course, this is only true for documents that have already been annotated by humans; the automatic pre-annotation usually fails to pick correct non-literal senses, as shown for a German idiom in (7). Furthermore, not much attention has so far been given to light verb constructions and non-decomposable idioms. As a result, most sentences containing such constructions do not have a gold annotation in the PMB yet, but only an automatically generated (i.e., bronze status) and semantically inadequate annotation using a literal sense of each word. Examples of this are shown in (8) and (3).

(4)

x_1, e_1, t_1
wedding.n.01(x_1), take_place.v.01(e_1), Theme(e_1, x_1), Time(e_1, t_1), DayOfWeek(t_1, saturday)

DRS for English PMB sentence 01/2506 *The wedding will take place on Saturday* (gold).

(5)

¬	s_1 ashamed.a.01(s_1), Experiencer(s_1, speaker)

DRS for German PMB sentence 03/2800 *Ich schäme mich nicht* "I'm not ashamed" (gold).

(6)

¬	x_1, e_1 spill.v.05(e_1), Agent(e_1, hearer), Theme(e_1, x_1), secret.n.01(x_1)

DRS for English PMB sentence 11/0958 *Don't spill the beans* (gold).

(7)

x_3, x_4, s_1
Order(x_3, "inneren"), Role(x_3, x_4), person.n.01(x_3), schweinehund.n.01(x_4), Patient(s_1, x_3), besiegen.a.01(s_1)

Partial DRS for German PMB sentence 17/1163 *den inneren Schweinehund zu besiegen* "to overcome one's weaker self" (not gold).

(8)

x_1, e_1
take.v.01(e_1), Agent(e_1, speaker), Theme(e_1, x_1), bath.n.02(x_1)

DRS for English PMB sentence 58/2404 *I'm taking a bath* (not gold).

In this work, we aim to improve the coverage of idioms in the PMB. This requires creating annotation guidelines that capture the semantics of such cases adequately while still fitting in with the lexicalized annotation framework of the PMB. It furthermore requires looking for idiom instances in the PMB and targeting them for annotation.

4. Extraction

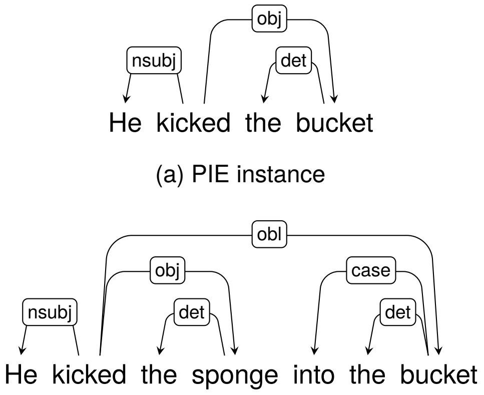
The first step was to find potential candidates for the annotation, i.e. sentences that contained German VID instances. To this end, we collected VID types from the *Redensarten-Index*[3] (transl. *Proverb-Index*), an electronic, privately maintained dictionary, which, contrary to the name, not only contains German proverbs but also an even larger number of idioms. At the time of this writing, the database comprises 15,661 entries. Since a lot of entries consist of several variants of the same expression, this number rises to 54,936 when counting every variant as a different type. After filtering out all the non-verbal expressions using parsing, 39,521 verbal ones remained.

After compiling a list of VID types, the next step was to find sentences in the PMB that contained instances of those VID types. We employed the parsing-based extraction method described in Haagsma (2020). This method only extracts sentences that contain the lemmata in the same dependency relations as the VID type, thus the focus of this approach is to increase precision by not extracting sentences that coincidentally comprise the same lemmata. Figure 1 shows two sentences that contain the tokens *kicked*, *the* and *bucket*, but only in (a) they have the desired dependency relations: NSUBJ between *bucket* and *kick* and OBJ between *kick* and *bucket*. In (b), the relation that holds between *kick* and *bucket* is OBL (for *oblique*) and accordingly the sentence would not be extracted, since it does not contain an instance of *kick the bucket* but only an accidental co-occurrence.

We employed UDPipe 2.12[4] (Straka, 2018) to

[3] https://www.redensarten-index.de/suche.php

[4] More specifically, the German model *german-gsd-ud-2.12-230717*



(a) PIE instance

(b) Not a PIE instance

Figure 1: Parsing-based extraction.

parse the gold, silver and bronze sentences of the German part of the PMB and subsequently used the method described above to extract sentences with VID candidates. This resulted in 6,187 sentences being extracted which were then prepared for annotation.

During this process not only instances of VID types were extracted, but also instances of their literal counterparts:

(9) Beth wurde von ihrem faulen Freund gefragt,
Beth was by her lazy friend asked,
ob sie **seine Hausaufgaben** für Geschichte
if she his homework for history
machen würde.
do would.
'Beth was asked by her lazy friend if she would do his homework for history.

In (9) we have an instance of *seine Hausaufgaben machen* (*to do one's homework*), but since it is the literal reading of this expression, we do not have an instance of the VID type (which means 'to prepare oneself'). These kind of literal instances are not relevant to the annotation of the PMB[5], but we decided to label them anyway in order to create a dataset of potentially idiomatic expressions (PIEs) as a byproduct. The term PIE encompasses both the literal and idiomatic meaning of an expression, thus we will use it from here on out when we talk about both at the same time.

5. Annotation

The annotation was conducted by three linguistically trained native speakers, with every sentence being annotated twice. Annotators were given text files where each instance to annotate came with a "form" with several questions they had to work through step by step (cf. Fig. 2).

In a first step, the guidelines were written and subsequently revised after a trial annotation of 50 sentences. However, due to the complex nature of the task, the guidelines kept on being revised multiple times throughout the whole process. To ensure consistency there was a subsequent correction step where every annotator revised their work once again. Weekly meetings with annotators were conducted throughout to discuss difficult cases and clarify the annotation guidelines.

The annotation consisted of several objectives:

1. Filter out false positives
2. Annotate the degree of idiomaticity
3. Judging the (non-)decomposability
4. Sense and role annotation

We will discuss these steps in more detail in the following.

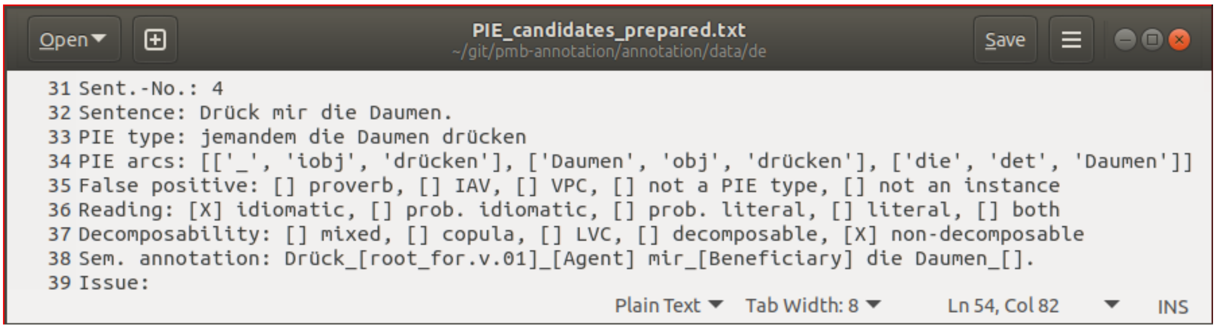
Firstly, due to errors during the extraction and the fact that we did not filter the list of idiomatic expressions other than for verbal types[6], there was a large number of false positives, i.e. types of expressions not of interest to us. Our focus was exclusively on what can be considered verbal idioms (VIDs) or, in rarer instances, light-verb constructions according to the PARSEME annotation guidelines 1.2, so verb senses that are only considered "multiword" because they obligatorily occur with a certain function word were to be ignored. These include verb-particle constructions (VPCs, e.g. *jmdm. etwas **an**tun* 'do something to somebody'), and inherently adpositional verbs (IAVs, e.g. ***zu** jmdm. halten* 'stand by sb.'). As we have seen in Section 3, VPCs are already handled satisfactorily in the PMB, and likewise IAVs, where the adposition is treated as part of the argument and does not contribute a sense on its own. Furthermore, proverbs were also not considered as these do not have free argument slots, contrary to idioms (e.g. *A watched pot never boils.*).

In the next step, the annotators had to decide whether the PIE instance fell into one of the following categories: IDIOMATIC, PROBABLY IDIOMATIC, PROBABLY LITERAL, LITERAL or BOTH. We gave the annotators the possibility to express uncertainty with the qualifier *probably* in order to account for the fact that some sentences did not have enough context to allow for maximum certainty regarding the reading - even if the annotator happened to be rather sure[7]. The label BOTH was intended for

[5] Because they usually can be treated compositionally.

[6] Manually filtering a list of 39,521 expressions would have been too time consuming.

[7] For example, because a certain PIE type was known to have one predominant reading.


Open
PIE_candidates_prepared.txt
~/git/pmb-annotation/annotation/data/de
Save
Plain Text
Tab Width: 8
Ln 54, Col 82
INS

```
Sent.-No.: 4
Sentence: Drück mir die Daumen.
PIE type: jemandem die Daumen drücken
PIE arcs: [['_', 'iobj', 'drücken'], ['Daumen', 'obj', 'drücken'], ['die', 'det', 'Daumen']]
False positive: [] proverb, [] IAV, [] VPC, [] not a PIE type, [] not an instance
Reading: [X] idiomatic, [] prob. idiomatic, [] prob. literal, [] literal, [] both
Decomposability: [] mixed, [] copula, [] LVC, [] decomposable, [X] non-decomposable
Sem. annotation: Drück_[root_for.v.01]_[Agent] mir_[Beneficiary] die Daumen_[].
Issue:
```

Figure 2: Text-based annotation interface, showing the sentence *Drück mir die Daumen!*, lit.: 'Squeeze your thumbs for me!', fig.: 'Wish me luck!'

cases in which both readings (IDIOMATIC and LITERAL) are active at the same time.

After that, the goal was to judge the level of decomposability of the expression. Besides the obvious labels, DECOMPOSABLE and NON-DECOMPOSABLE, the annotators could also choose the labels LVC, COPULA and MIXED. The latter three categories will be discussed in the next section in greater detail.

Strictly speaking, the previous step was not really necessary, but served as a kind of priming for the last step: the semantic annotation of the idiom and its arguments. During this step, the annotators were supposed to choose the WordNet sense (Fellbaum, 1998) that most closely corresponded to the meaning of the idiom and add it to the sentence. In order to do this, the annotators had to decide on the level of decomposability anyway because the number of senses added to the VID depended on this. Consider the next two examples for illustration:

(10) Er_[Experiencer] **schwimmt**_[buck.v.02]
He swims
gegen_[] **den**_[]
against the
Strom_[Stimulus]_[trend.n.01].
tide.
'He bucks the trend.'

(11) **Stecke**_[despair.v.01]_[Experiencer] nicht
Bury not
den Kopf_[] **in den Sand**_[]!
the head in the sand!
'Don't despair!'

Example (10) shows an instance of the VID *gegen den Strom schwimmen* (*swim against the tide* ⇒ 'buck the trend'), which is decomposable as we can map the individual idiomatic meanings to the components: 'buck' → *swim* and 'trend' → *tide*. Consequently, the two WordNet senses *buck.v.02* and *trend.n.01* were added. The example furthermore shows that in addition to the senses we also added the semantic roles of the predicate's arguments, in this case *Experiencer* and *Stimulus*. Annotators were instructed to use WordNet Search 3.1[8] for finding senses, and VerbAtlas (Di Fabio et al., 2019) for mapping them to VerbNet-style rolesets, but to prefer PMB-specific conventions when in doubt. As can be seen, the senses were added by suffixing an underscore followed by brackets to a component. If a component was annotated with a sense and a semantic role, the latter always preceded the former (first *Stimulus* then *trend.n.01* in this case).

In example (11), on the other hand, we have an instance of the non-decomposable VID *den Kopf in den Sand stecken* (*to put the head in the sand* ⇒ 'to despair'). It is non-decomposable as it is not possible to decompose the overall idiomatic meaning into individual meanings. For non-decomposable VIDs the WordNet sense (*despair.v.01* in this case) was added to the verbal head of the expression, while the other brackets were left empty.

Apart from VIDs we also annotated for LVCs as they are also not handled in the desired manner in the PMB:

(12) Die Generation_[Theme] der
The generation of
Zeitzeugen **geht**_[end.v.01]
contemporary witnesses goes
zu_[] **Ende**_[] [...]
to end [...]
'The Generation of contemporary witnesses is ending.'

Example (12) contains an instance of the LVC *zu Ende gehen* (*to go to end* ⇒ 'to end'). We consider this a special case of non-decomposability since no part of the meaning could ever be mapped to the semantically bleached verbal part. To ensure consistency we nevertheless add the sense

[8] http://wordnetweb.princeton.edu/perl/webwn

(*end.v.01*) to the verbal part of the expression. Please note that we did not annotate for expressions that according to the PARSEME annotation guidelines would be considered LVC.cause, i.e. the verb indicates the cause of the event (e.g. *to grant rights* or *to provoke a reaction*).

6. Annotation Results and Discussion

6.1. Inter-annotator agreement

For computing agreement, we excluded 341 sentences that had been discussed in annotation meetings, thus had not been annotated by two annotators independently. For simplicity, we also excluded 18 sentences that for various reasons did not have exactly 2 annotations and 7 sentences where one or both annotators detected more than one instance of the same idiom.

On the remaining 5,821 sentences, we classified annotators' decisions both broadly into "idiom" or "not an idiom", and more finely by, e.g. decomposability class or false positive class. On the coarse-grained comparison, annotators agreed in 3,448 cases that something is not an idiom and should thus not receive a detailed semantic annotation. In 1,945 cases they agreed it is an idiom. And in 428 cases they disagreed on this. Coarse-grained agreement is strong (Cohen's $\kappa = .8433$).

On the fine-grained comparison, annotators agreed in 4,230 cases and disagreed in 1,591 cases, yielding a moderate $\kappa = .6311$. Table 1 shows how frequent each class is, looking only at instances where annotators agree. We can see that most instances extracted are false positives, in particular cases where the extracted structure is not an instance of the idiom type, as in Figure 1b. Among the instances unanimously classified as idioms, a large majority is annotated as non-decomposable.

Table 2 shows the ten most frequently disagreed upon classes. In many cases, annotators agree that the items are not relevant to our annotation goal, they just disagree on why (e.g., IAV vs. not an instance). In other cases, annotators came to different conclusions regarding decomposability. Finally, there are cases where one annotator annotated the item as a non-decomposable idiom whereas the other deemed it not an instance, an IAV, not a verbal PIE type, or literal.

For the sense and role annotation of items that both annotators classified as an idiom, we look at whether both annotators selected the same word as the syntactic head of the idiom (head selection), whether they assigned the selected head the same sense (head sense classification), and for each word in the sentence whether they marked it as the head of an argument that is part of the (decomposable) idiom (internal argument identification), or as an argument that is not part of the idiom (external argument identification). For unanimously identified internal arguments, we also look at role and sense classification, and for unanimously identified external arguments, at role classification. Table 3 shows the results, with strong agreement for head selection and argument identification, weak to moderate agreement for head sense classification, and moderate to strong agreement for argument role and sense classification scores.

not an idiom	**3,448**
not an instance	1,968
IAV	194
VPC	149
proverb	142
literal	121
not a verbal PIE type	90
idiom	**1,945**
non-decomposable	1,335
decomposable	186
LVC	24
copula	19
mixed	2

Table 1: Unanimously classified PIEs by frequency. Numbers in bold represent coarse-grained agreement.

IAV, not an instance	349
literal, not an instance	195
decomposable, non-decomposable	181
non-decomposable, not an instance	**136**
LVC, non-decomposable	108
not a verbal PIE type, not an instance	91
IAV, non-decomposable	**73**
non-decomposable, not a verbal PIE type	**43**
IAV, literal	41
literal, non-decomposable	**39**

Table 2: Most frequent disagreements in PIE classification. Entries in bold are not only fine-grained but also coarse-grained disagreement.

6.2. Challenges to the annotation

In the following we will discuss some of the reasons that made the task quite challenging. As mentioned above, the guidelines were revised multiple times during the annotation process.

Decomposability One of these revisions consisted of adding another category w.r.t. decomposability. During the annotation it became clear that some expressions do not fit the binary distinction of decomposability presented above:

Head selection	.9769
Head sense classification	.5862
Internal argument identification	.9914
Internal argument role classification	.7296
Internal argument sense classification	.6824
External argument identification	.9845
External argument role classification	.8352

Table 3: Agreement scores for semantic annotation of idioms. Head selection is given in terms of raw agreement; the other scores are Cohen's κ scores.

(13) Tom_[Agent] **legte**_[reveal.v.02] **die**_[]
Tom laid the
Karten_[Topic]_[intention.n.01] **auf**_[]
cards on
den_[] **Tisch**_[].
the table.
'Tom revealed his intentions'.

Example (13) shows an instance of the VID *die Karten auf den Tisch legen* (*to lay the cards on the table* ⇒ 'to reveal one's intentions'). It is decomposable in the sense that we can map 'reveal' to *auf den Tisch legen* and 'intentions' to *Karten*, but there is no part of the meaning we can map to *Tisch* individually, i.e *auf den Tisch legen* itself is non-decomposable. To accomodate for these kind of instances, we added the category MIXED to the possible choices for decomposability.

Another frequently discussed question was whether to prioritize decomposition even when a non-decomposable analysis would have been more convenient because a very suitable sense was available:

(14) Der Gouverneur_[Agent] **setzte**_[set.v.05]
The governor set
die Häftlinge_[Patient] **auf freien**
the prisoners on free
Fuß_[Result]_[free.a.01].
foot.
'The governor set the prisoners free'.

Example (14) contains an instance of the VID *jmdn. auf freien Fuß setzen* (*to set sb. on free foot* ⇒ 'to set sb. free'), so the WordNet sense *set_free.v.01* would have been very fitting, but since we decided to prioritize the decomposition of the expression in such cases we opted for a decomposable analysis which seems less elegant.

Missing senses As one can imagine, it is not always straightforward to map a German idiom to an English WordNet sense. Sometimes there are two or more equally plausible possibilities, leading to spurious disagreement, e.g. *dazzle.v.02* or *stagger.v.04* for *jmdm. den Atem rauben* 'to take sb.'s breath away'. In case of missing verbal synsets, we were often able to use a nominal, adjectival, or adverbial one instead, as in (15).

(15) Dichter_[AttributeOf] wie Milton
Poets like Milton
sind_[rare.a.03] **dünn**_[] **gesät**_[].
are thinly sowed.
'Poets like Milton are few and far between.'

But sometimes we were hardly able to find any fitting sense at all.

(16) Tom **hat nichts zu verlieren**.
Tom has nothing to lose.
'Tom has nothing to lose.'

For example, the expression *nichts zu verlieren haben* 'to have nothing to lose' means something along the lines of being desperate and prone to dangerous behavior, but we were not able to find a synset capturing this, as, e.g. *desperate.a.03* seemed both too general and too specific, so we did not annotate (16), although in cases were we found a synset that was a bit too general but not too specific we usually accepted it, as in (17).

(17) er_[Agent] **gab**_[give.v.20] **ihm**_[Patient]
he gave him
einen tüchtigen Fußtritt_[Theme] **mit**_[]
a hearty kick with
auf_[] **den**_[] **Weg**_[]
on the way
'he gave him a good kick (as he was leaving)'

Some idioms have an emphatic meaning component not captured by the synset we assigned it, as in (18).

(18) Tom_[AttributeOf] **schwimmt**_[rich.a.01]
Tom swims
im Geld_[].
in the money.
'Tom is rolling in money.'

As a last resort when unable to find a roughly fitting synset, we would create a new one:

(19) Mir_[Experiencer] **fällt**_[cabin_fever.n.00]
Me falls
die_[] **Decke**_[] **auf**_[] **den**_[] **Kopf**_[].
the ceiling on the head.
'I'm starting to get cabin fever'.

The expression *jmdm. fällt die Decke auf den Kopf* (*the ceiling falls on sb's head*) alludes to the negative psychological effects someone can experience when confined to a small space for a long period of time. In English, the term *cabin fever* exists to describe this state, but it is not available in WordNet. And neither is any equivalent sense, so in such cases, we made a sense up which we suffixed with 00 (*cabin_fever.n.00* in (19)).

Collocations Lastly, the status of collocations was discussed frequently. Although we were not aware of it during annotation, we find the distinction between *idioms of encoding* and *idioms of decoding* (Fillmore et al., 1988; Richter and Sailer, 2014) helpful. Idioms of decoding are idioms proper: a listener has to know the expression to understand it, e.g. *ins Gras beißen*, lit. 'bite into the grass', 'kick the bucket'. Idioms of encoding require the speaker to know an expression to encode the meaning idiomatically, e.g. to know to say *Zähne putzen*, lit. 'clean teeth', 'brush teeth', and not *Zähne sauber machen*, lit. 'make teeth clean', although both encode the meaning compositionally and are understandable without having the expression in the mental lexicon. Mere idioms of encoding are sometimes called collocations, and were out of scope for this annotation project. But sometimes the difference is hard to tell.

(20) Endlich **zeigte** er **sein wahres Gesicht**.
Finally shows he his true face.
'Finally he reveals his real personality.'

(21) Wir sollten das wohl **unter vier**
We should that probably among four
Augen besprechen.
eyes talk about.
'We should probably discuss this in private.'

For example, in (20), one can argue that *sein wahres Gesicht zeigen* is an idiom of decoding because *Gesicht* with the sense *personality* is not often, perhaps never found outside of this expression, whereas *zeigen* with the sense *reveal* is quite common. Another example is shown in (21), where one can likewise argue that the adverbial phrase *unter vier Augen* in the sense *in private* usually only occurs with the verb *besprechen* or a small set of near-synonyms like *bereden*, *diskutieren*. We did not annotate these examples in the end and leave defining a sharper criterion for distinguishing idioms from collocations for future work.

7. Conclusions and Future Work

Idioms present many challenges to semantic annotation in a lexicalized framework, which leads to them being underrepresented or inadequately annotated in sembanks. In this work, we have carried out a targeted annotation of German idioms in the Parallel Meaning Bank by automatically detecting instances of potentially idiomatic expressions (PIEs) and annotating them for their idiomatic status, as well as their semantics, including WordNet senses and VerbNet semantic roles. Many automatically detected PIEs were false positives; of the rest, most received non-decomposable analyses, some decomposable ones, and some received special labels like MIXED, COPULA, or LVC. Inter-annotator agreement across the subtasks is very encouraging considering the complexity of the task, with the lowest score achieved for word sense disambiguation, unsurprising given that English synsets are many and there is often no unique mapping from German idioms and words to them. As our qualitative analysis of the results shows, there are also many subtle difficulties in classifying PIEs.

The next challenge will be to actually integrate the produced annotations into the PMB so as to get closer to a gold standard semantic annotation for sentences containing idioms. We are preparing a translation of the annotations into *bits of wisdom*, the format in which human annotator decisions are stored in the PMB and then inserted into the PMB's dynamic annotation workflow. Assigning senses and roles is relatively straightforward; however, for non-decomposable idioms, we also have to make sure that the arguments get assigned λ-DRSs that do not contribute concepts, which will require adding some new rules to Boxer, the rule-based component computing meaning representations based on syntax and token-level annotations. The documents receiving the annotations will automatically receive silver status and have to be checked manually again to receive gold status. This will make the PMB a more comprehensive and challenging testbed for data-driven DRS parsers such as van Noord et al. (2020) or Shen and Evang (2022), whose ability to handle idioms future work will also address. Furthermore, an analogous annotation project is currently underway for English idoms in the PMB.

Acknowledgments

We would like to thank the anonymous reviewers for their feedback. We would also like to thank our annotators for their work. This work was carried out in the MWE-SemPrE project funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 467699802.

8. Bibliographical References

Julia Bonn, Andrew Cowell, Jan Hajič, Alexis Palmer, Martha Palmer, James Pustejovsky, Haibo Sun, Zdenka Uresova, Shira Wein, Nianwen Xue, and Jin Zhao. 2023. UMR annotation of multiword expressions. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 99–109, Nancy, France. Association for Computational Linguistics.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language*, 64:501–538.

Hessel Haagsma. 2020. *A Bigger Fish to Fry: Scaling up the Automatic Understanding of Idiomatic Expressions*. Ph.D. thesis, Rijksuniversiteit Groningen.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Springer, Dordrecht.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538. Publisher: Linguistic Society of America.

Frank Richter and Manfred Sailer. 2014. Idiome mit phraseologisierten Teilsätzen: eine Fallstudie zur Formalisierung von Konstruktionen im Rahmen der HPSG. In Alexander Lasch and Alexander Ziem, editors, *Grammatik als Netzwerk von Konstruktionen: Sprachwissen im Fokus der Konstruktionsgrammatik*, pages 291–312. De Gruyter, Berlin, Boston.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, and Gyri Smørdal Losnegaard. 2015. PARSEME–PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*.

Minxing Shen and Kilian Evang. 2022. DRS parsing as sequence labeling. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 213–225, Seattle, Washington. Association for Computational Linguistics.

Mark Steedman. 2001. *The Syntactic Process*. The MIT Press.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Jens E. L. van Gysel, Jayeol Chun Meagan Vigus, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmerand James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *Künstliche Intelligenz*, 35:343–360.

Rik van Noord, Antonio Toral, and Johan Bos. 2020. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

9. Language Resource References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Lasha Abzianidze, Rik van Noord, Chunliu Wang, and Johan Bos. 2020. The parallel meaning bank: A framework for semantically annotating multiple languages. *Applied mathematics and informatics*, 25(2):45–60.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Karin Kipper Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO redwoods treebank: Motivation and preliminary applications. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.

Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Verginica Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, and Voula Giouli. 2018. Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, and Voula Giouli. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118.

Agata Savary, Chérifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, and Sara Stymne. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547.

Universal Feature-based Morphological Trees

Federica Gamba, Abishek Stephen, Zdeněk Žabokrtský
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
{gamba, stephen, zabokrtsky}@ufal.mff.cuni.cz

Abstract

The paper proposes a novel data representation inspired by Universal Dependencies (UD) syntactic trees, which are extended to capture the internal morphological structure of word forms. As a result, morphological segmentation is incorporated within the UD representation of syntactic dependencies. To derive the proposed data structure we leverage existing annotation of UD treebanks as well as available resources for segmentation, and we select 10 languages to work with in the presented case study. Additionally, statistical analysis reveals a robust correlation between morphs and sets of morphological features of words. We thus align the morphs to the observed feature inventories capturing the morphological meaning of morphs. Through the beneficial exploitation of cross-lingual correspondence of morphs, the proposed syntactic representation based on morphological segmentation proves to enhance the comparability of sentence structures across languages.

Keywords: Morphs, Universal Segmentations, Universal Dependencies

1. Introduction

Universal Dependencies (UD) (de Marneffe et al., 2021) is a framework for consistent annotation of natural language data across languages. The UD project develops cross-linguistically consistent treebanks to facilitate multilingual and cross-lingual parsing research from a typological perspective.[1] However, the syntactic annotation proposed by UD, along with the standard tokenization often based on white-space,[2] poses some challenges to actual comparability across languages, as different languages may adopt different strategies to express the same phenomenon. Consider, for instance, the English sentence *I will go through a forest*, translatable in Czech as *Půjdu lesem*.

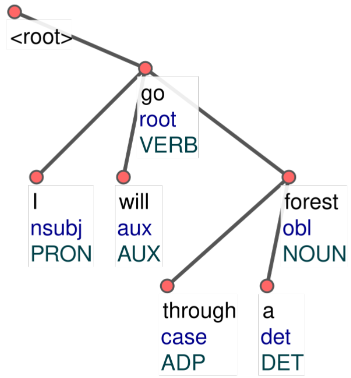


Figure 1: UD tree for the English sentence *I will go through a forest.*

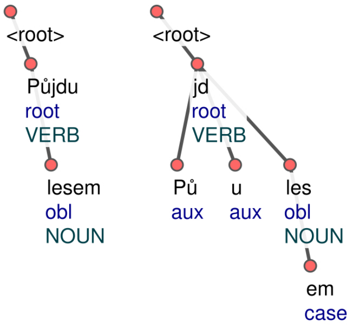


Figure 2: UD and morphological tree for the Czech sentence *Půjdu lesem. Pů* – a prefix expressing future tense, *jd* – the room morph for 'to go', *u* – a 1st pers. sg. conjugation ending, *les* – the root morph for 'forest', *em* – instr. sg. masc. declination ending.

These two equivalent sentences exhibit noticeable differences already in the token count, and their dissimilarity is reflected in their respective dependency tree structures. Nonetheless, a closer look at the sentences reveals that splitting word forms based on their morphological segmentation leads to a better mapping concerning isomorphy of trees and alignment of nodes,[3] allowing for greater comparability. Notably, in this example, Czech encodes future tense through the prefix *pů*, whereas the ending *em* for instrumental case in *lesem* expresses movement through (Figure 1, 2). Similarly, at the surface level the German compound *Finanzkrise* 'financial crisis' does not correspond –

[1] https://universaldependencies.org/.

[2] At least in the case of languages with the alphabetic writing system.

[3] At the word level, we observe a 3:1, 3:1 node alignment; at the morph level, node alignment is 1:1, 1:1, 1:1, 1:1, 1:0 (article unexpressed in Czech), 1:1.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 125–137. May 25, 2024.

in terms of structure and token count – to its Czech counterpart *finanční krize*. However, if we segment the two members that led to the formation of the compound (*Finanzen* + *Krise*), we obtain a clear correspondence of the German and Czech forms. A syntactic representation based on morphological segmentation could thus enhance the cross-lingual comparability of languages that e.g. exhibit different amounts of inflection or productivity in compounding.

Additionally, what emerges from the observation of segmented morphs[4] is that morphological features often tend to be associated to specific morphs. For instance, in the English word *letters* the morph *s* can be morphologically interpreted as an encoding for plurality. The morphological specification of a (syntactic) word form is encoded by a set of features in UD representing the lexical and grammatical properties. UD differentiates between lexical and inflectional features, where the former are an attribute of lemmas and the latter of word forms. This approach is convenient and productive in capturing the morphosyntactic functions of word forms, which fits the goal of UD, but it will not be incorrect to postulate that such lexical or grammatical functions can be encapsulated within morphs in a word form.

Thus, this study aims to propose a novel data representation, which exploits UD-like trees to represent simultaneously the UD-like syntactic sentence representation as well as the internal structure of word forms (hence taking the Item-and-Arrangement perspective on morphology (Bram, 2012)), which is merged within a single dependency tree. Using the inventory of universal morphological features in UD, we also investigate whether a strong correlation can be found between a given morph and a feature value, and then align the morphs to the observed feature that captures the lexical and grammatical functions of morphs. We thus propose a data structure that intertwines syntax and morphology with the goal of increasing comparability across languages.

The remainder of the paper is structured as follows. In Section 2 we present the related work, while Section 3 offers an overview of the resources that we employ for the present study. Section 4 details how such resources are exploited, focusing on the manipulation of treebank nodes and feature extraction, as well as discussing the strategy devised to comply with the UD schema. Section 5 shows the UD-like morphological trees that result from the present work, while Section 6 concludes the paper and outlines future research directions.

2. Related Work

The idea of representing the internal structure of words has been previously explored, especially for non-alphabetic languages such as Chinese. In these kinds of languages, the issue of delimiting word boundaries is far from trivial and requires alternative strategies to be inspected. For instance, Zhao (2009) investigates internal character dependencies inside a word as a result of the attempt to handle word boundaries by identifying character-level dependencies.

Li (2011) elaborates on this approach by suggesting to recover word structures in morphological analysis. One of the reasons for this lies in the observation that there exist many different annotation standards for Chinese word segmentation, which could even cause inconsistency in the same corpus.[5] As we are working with alphabetical languages, their motivation for the work differs from ours. Additionally, we adopt dependency structures, while they work with constituency trees.

Concrete applications in the parsing of the approach in Li (2011) are described e.g. by Zhang et al. (2013), who annotate internal structures of words and then build a joint segmentation, part-of-speech (POS) tagging and phrase-structure parsing system. Zhang et al. (2014) integrate inter-word syntactic dependencies and intra-word dependencies, differentiating intra- and inter-word dependencies by the arc type to achieve results comparable to conventional resources.

In the case of languages with alphabetical writing systems, CELEX (Baayen et al., 1995) represents morphological word structure for Dutch, English, and German in the shape of a tree. Steiner (2017), e.g., exploits the resource in combination with GermaNet (Hamp and Feldweg, 1997). Morphological and compound information is extracted from the two resources respectively, and reused to build a so-called morphological treebank for German. However, such a morphological treebank consists of tree-shaped single tree-words only, without including any kind of syntactic information at a sentence level.

An example of integration of morphology and syntax is provided by the UD treebank for Beja (Kahane et al., 2021), a Cushitic language spoken in Sudan. In the treebank, a morph-based tokenization instead of a word-based one is adopted. All affixes are dependent on the stem and are assigned UD deprels corresponding to their functional role, with an additional `:aff` subtype (e.g., subject pronominal affixes are marked as `nsubj:aff`).

[4] Due to the ambiguous usage of the term 'morphemes', we use the term 'morphs' henceforth based on Haspelmath (2020).

[5] For instance, *vice president* could be considered as a single word or split into two words.

3. Exploited Resources

For the present study, we exploit the resources described hereafter. UniSegments, UniMorph, and SIGMORPHON data are selected to obtain the segmentation, which we employ to manipulate UD trees. The selection of the languages primarily stems from their availability across all resources.[6]

UniSegments UniSegments (Žabokrtský et al., 2022) is a collection of harmonized versions of selected resources relevant for segmentation, whose data have been converted to a common scheme. It comprises 17 existing data resources featuring information about segmentation in 32 languages. The level of granularity of information varies across the different resources. Some of them classify segments specifying whether they are either roots, prefixes/suffixes, inflectional endings, or zero morph(eme)s; yet, despite using the same labels, they adopt different definitions of the classes. In the attempt to devise a truly shared schema, the creators of UniSegments chose to preserve the parts that require deep in-language expertise (e.g., lemmas), unify the information available in most resources (POS tags and, to some extent, segmentation), and keep as much of the language/resource-specific information as possible unchanged (Žabokrtský et al., 2022). This ensures a balance between the diverse levels of granularity observed in the resources but does not guarantee their full conformity. Inevitably, such discrepancies among the resources will be indirectly reflected in our data. At times, UniSegments includes more than one resource for the same language; in such cases, we select only one resource. We work with DeriNet (Vidra et al., 2021) for Czech, MorphoLex (Sánchez-Gutiérrez et al., 2018) for English, Demonette (Hathout and Namer, 2014) for French, DerIvaTario (Talamo et al., 2016) for Italian, WordFormationLatin (Litta et al., 2016) for Latin, and MorphyNet (Batsuren et al., 2021) for Catalan, Finnish, German, Hungarian, and Portuguese.

UniMorph The Universal Morphology (UniMorph) (McCarthy et al., 2020) project aims at providing instantiated normalized morphological paradigms for hundreds of diverse world languages, provided in a shared morphological schema. As far as the languages we include in our work are concerned, morphological information is extracted from Wiktionary (e.g., for Finnish) or derived from existing morphological dictionaries which are publicly hosted on the LINDAT/CLARIAH-CZ repository (for English, French, German, Italian).[7] Since information about vowel length is available for Latin data in UniMorph, data normalization is needed before undertaking the manipulation of nodes in treebanks.[8]

SIGMORPHON Some datasets were made available for the SIGMORPHON 2022 Shared Task on Morpheme Segmentation (Batsuren et al., 2022). We choose to exploit Czech gold annotated data, as the quality of the results could prove to be positively affected.

Universal Dependencies A brief introduction to UD is available in Section 1. For the languages under study, we select the following treebanks from version 2.12 (Zeman, 2023). Whenever a Parallel Universal Dependencies (PUD) treebank (Zeman et al., 2017) is available we include it, as the PUD collection can provide interesting insights in terms of parallel, cross-lingual comparison. Additionally, we also select PDT (Hajič et al., 2020) for Czech, GUM (Zeldes, 2017) for English, TDT (Pyysalo et al., 2015) for Finnish, GSD (McDonald et al., 2013) for French and German, ISDT (Bosco et al., 2013) for Italian, and Bosque (Rademaker et al., 2017) for Portuguese. We employ AnCora (Taulé et al., 2008) for Catalan, Szeged (Vincze et al., 2010; Vincze et al., 2017) for Hungarian, and ITTB (Passarotti, 2019) for Latin, for which no PUD treebank is available.

4. Workflow

We now describe the strategy designed to process the selected data and extract from it all the exploitable information. It mainly revolves around two main tasks: on the one hand, the manipulation of nodes in treebanks based on the segmentation contained in the selected sources (Subsection 4.1); on the other hand, the process of alignment between universal features and morphs (Subsection 4.2). As a result, we release a set of treebanks where morphological segmentation is incorporated within the UD representation of syntactic dependencies.[9] How the morphs are integrated into the UD annotation is discussed in Subsection 4.3.

4.1. Manipulation of Treebank Nodes

As a first step, we convert the official UD treebanks to morphologically segmented treebanks, as described hereafter and illustrated in Figure 3.

To manipulate data we exploit Udapi (Popel et al., 2017), a framework providing an application programming interface for UD data. The code that performs the transformation is not language-specific,

[6] With the only exception of SIGMORPHON.

[7] Additionally, for some low-resource languages and dialects the data mainly comes from linguists who study them. Data augmentation in a semi-supervised way was also experimented with for Tagalog.

[8] For instance, *ă* and *ā* are normalized as *a*.

[9] Both the code and the set of treebanks are openly available at https://github.com/fjambe/feature-based-morpho-trees/.

provided that resources featuring morphological information (e.g., about segmentation, derivation, inflection) are available. It takes as input the UD treebank to manipulate and outputs a version of it where morphological trees of segmented words are blended in UD tree-shaped sentence representation, within a well-formed CoNLL-U file.

By iterating over each node, we check whether information about morphological segmentation of the node form or lemma (as further explained later) is available in any of the exploited resources, i.e. UniSegments and UniMorph mainly, as well as SIGMORPHON gold data for Czech.[10]

Step 0: SIGMORPHON data. In the case of Czech, we exploit SIGMORPHON manually annotated data as an additional resource. As a preliminary step in the workflow, for each form we first check whether it occurs in SIGMORPHON data; if it does, we split the form according to this segmentation. Since SIGMORPHON data only provides splitting, with no additional information about the resulting morphs, deciding which morph of the word should be considered the root is not straightforward. Thus, we decide to select as root the least frequent morph among those we identify within the word. Morph frequencies were calculated initially on the whole dataset. Whenever a form is found in SIGMORPHON, we then cease looking for possible additional segmentations, since forms in SIGMORPHON data are fully segmented. If, conversely, the word form is not retrieved at this stage, we continue with the procedure valid for all languages.

Step 1: segmented lemma. The first step consists of looking up the word lemma in UniSegments. If a match is found and a segmentation is available for the retrieved lemma,[11] the information just retrieved is now stored, to be exploited subsequently to segment the node. For instance, the Czech word *prokonzul* ‘proconsul’ is found in UniSegments as well as provided with a segmentation (*pro* + *konzul*).

Step 2: (un)inflected form, segmented lemma. Afterward, we check if the node form corresponds to its lemma, i.e. if the token is not an inflected form. If this is the case, we proceed to split the form based on the segmentation retrieved in UniSegments, as illustrated by the *prokonzul* example. Conversely, if the form is inflected we postpone the splitting until we have gathered more information about the word ending. For this purpose, we begin by verifying whether the form is listed in UniMorph, which comprises a catalog of inflected forms. If this proves to be the circumstance, we combine the information from UniSegments with information about inflection retrieved from UniMorph. See e.g. the Catalan plural form *culturals* ‘cultural’, whose lemma is split in UniSegments as *cultur* + *al*, while UniMorph provides the morph *-s* for plural. If, conversely, no match is found in UniMorph, we design a strategy to obtain an approximation of the inflectional ending by comparing character by character the two strings (form and lemma) and extract as ending the substring starting after the last shared character and extending till the end of the word form. It is the case of the English verb form *shortened*, split as *short* + *en* in UniSegments, and for which we extract the ending *-ed*.

Step 3: inflected form, unsegmented lemma. If the node lemma is not found in UniSegments, we inspect whether the node form occurs in UniMorph only. If it does, we extract the information from UniMorph and proceed to segment at least the inflectional ending of the word, as in the case of the French *travaillait* ‘(s)he worked’, third person singular form of the imperfect tense of the verb *travailler* ‘to work’. The form is segmented in UniMorph as *travailler* (lemma) + *ait* (ending).

Step 4: uninflected form, unsegmented lemma. In case the word is not comprised in either UniSegments or in UniMorph, i.e. if the node lemma and the node form do not represent entries of either of the two resources respectively, we do not implement any morphological splitting of the node and we proceed to the next one. That is, for instance, what happens with the Latin form *caelum* ‘sky’, corresponding to nominative, accusative, and vocative singular. Since for Latin nouns the nominative singular form is chosen as lemma, the form is not split in UniMorph; given that it is not segmented in UniSegments either, no morphological splitting can be performed on such form.

Practically, in the CoNLL-U file we handle morphologically segmented words as UD multi-word tokens (MWTs). Yet, such a decision may generate ambiguity, as it could be complex to distinguish original MWTs from morphological MWTs,[12] especially when they occur jointly (i.e., a MWT which we split further). Therefore, we decide to signal

[10] At this moment, we search only for a single best segmentation for each node, without handling possible ambiguities. Considering multiple segmentations may turns out to be necessary, especially in heavily ambiguous languages such as Arabic; morphological lattices (More et al., 2018) could be then useful for representing sets of alternative segmentations.

[11] Some of the lemmas included in UniSegments are not provided with a segmentation. See, for instance, Czech words *rok* ‘year’ or *jazyk* ‘language’, for which the only segment identified is the root, spanning over the whole word.

[12] Within the expression ‘morphological MWT’ we intend to use ‘MWT’ only in the technical sense of the UD label.

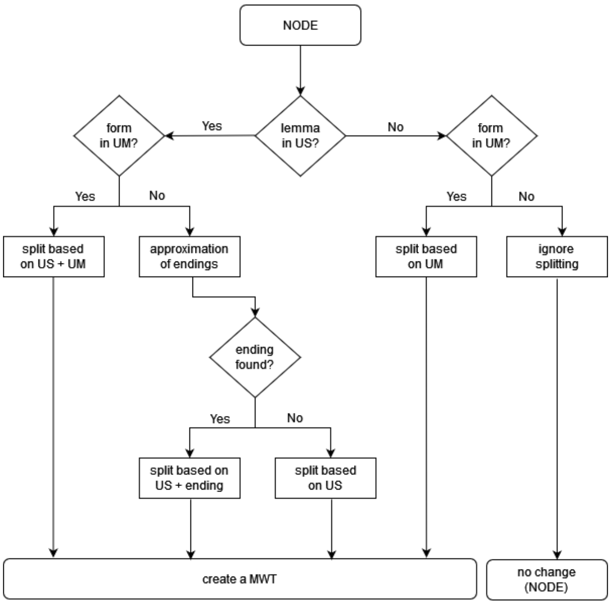


Figure 3: Flowchart of the node manipulation process (US: UniSegments, UM: UniMorph).

morphology-based split elements of MWTs through the deprel subtype `:morph` (see Subsection 4.3).

Since we are proposing a novel data representation, we have no gold data to rely on to assess the quality of the output of our algorithm. In light of this, we created a random sample of 20 French words and segmented them manually, which resulted in identifying 56 morphs. Of the 56 morphs in this gold data, 8 (14%) were correctly identified by UniSegments[13] alone, 18 (32%) by UniMorph alone, and 27 (48%) by our algorithm. Even though this sample is very small, it can be argued that combining the resources using our algorithm leads to a considerable improvement in the segmentation quality.

4.2. Feature Extraction

Additionally, by exploiting the statistical measures described hereafter we investigate whether and how morphs and UD feature sets align, to assess if specific feature inventories somehow capture the morphological meaning of morphs.

Similarly to what was done for node manipulation, we exploit the information contained in segmentation resources (in this case, UniSegments only) and in UD treebanks. Specifically, if a word form occurs in the treebank under study, and its lemma is also present in UniSegments, we segment it based on the segmentation provided by UniSegments.[14] For example, in Catalan the word *estacional* 'seasonal' is present in the UD Catalan AnCora treebank and also in UniSegments, following which it is split as *estacion* and *al*.[15] After having obtained the segmentations of the word forms from UniSegments, the UD feature set that is originally attributed to the word form is associated to the individual morphs the word form has been split into. For instance, the Hungarian word *gyerekek* 'children' in the UD Hungarian-Szeged treebank has the feature set `Number=Plur|Case=Nom`. Based on the segmentation data for Hungarian in UniSegments, the word form is split as *gyerek* + *ek*; we assign the original feature set to both *gyerek* and *ek*. In the following step, the feature set is split into individual features and is assigned to the morphs. As a result, we now have two instances of *gyerek*, one with feature `Number=Plur` and the other with feature `Case=Nom`; the same applies to *ek*. In this manner, for every possible feature, we create an inventory of morphs to which the feature has been associated. For each feature-morph pair we calculate the joint frequency of locating a morph given a feature and the ΔP scores (Jenkins and Ward, 1965). According to Schneider (2020), ΔP is a measure of cue validity, i.e. it measures how strongly two events are linked.[16] ΔP can be thus used to calculate collocation strength. Since it is a unidirectional dependency measure it can be decomposed in two distinct formulae, one for the forward-directed ΔP and the other for the backward counterpart. Using ΔP, we obtain the measure of the strength of correspondence between a morph and a feature, and vice versa. It is reasonable to use such a unidirectional measure because the association of a morph and a feature is asymmetric. The ΔP scores are between -1 and 1.

$$\Delta P_{forward} = P(m|f) - P(m|\neg f) \tag{1}$$

$$\Delta P_{backward} = P(f|m) - P(f|\neg m) \tag{2}$$

In equations (1) and (2), *m* stands for morph and *f* stands for feature. P(m|f) is the conditional probability of locating a morph given a feature among the other conditional probabilities in the equations. In Table 1, we present the ΔP forward and the ΔP backward scores for the morph *ing* in English given

[13] Specifically, among the available resources for French we selected Demonette.

[14] All the steps described in this paragraph are not applied to the same files employed for manipulation of treebank nodes.

[15] The example points out how morphological segmentation still presents several open issues. *Estacional* could probably be split further, by identifying *st* as the true core of the word, and *(c)ion* as another affix. We do not address the choices made in terms of segmentation, and work with the resources in their current state, however being aware that possible alternative segmentations could be proposed.

[16] The question how to extract combinations of features (conjunctions and disjunctions), which is relevant especially with inflectional affixes, is left for future research.

different morphological features. We find that the morph *ing* has the strongest relation with the feature VerbForm=Ger. What this indicates is the fact that the VerbForm=Ger strongly correlates to the morph *ing* as indicated by ΔP forward; the ΔP backward scores show the potential feature attributes like Tense=Pres, VerbForm=Part as well as the highest ranked feature VerbForm=Ger. Hence by comparing the ΔP forward and backward scores some signals could be extracted for morph and feature correspondences. While for a well-resourced language like English, such findings are not surprising, interesting correspondences could emerge in the case of less described languages.

Morph	Feature	ΔP forward	ΔP backward
ing	Degree=Pos	-0.058	-0.118
ing	Number=Sing	-0.090	-0.287
ing	Number=Plur	-0.091	-0.201
ing	Mood=Ind	-0.096	-0.144
ing	Person=3	-0.094	-0.127
ing	Tense=Pres	0.139	0.148
ing	VerbForm=Fin	-0.097	-0.152
ing	VerbForm=Part	0.120	0.136
ing	**VerbForm=Ger**	**0.966**	**0.710**
ing	Polarity=Neg	-0.004	-0.001

Table 1: Probabilities of the morph *ing* in English.

In Table 2, we observe that the morph *ung* in German has the highest ΔP scores for the feature Gender=Fem. The association with other features is due to the co-occurrence with other morphs in a word form. For example, the feature set for the German word *Kleidung* 'clothing' is Case=Nom|Gender=Fem|Number=Sing. The observed co-occurrences with other features can be explained by the allocation of the original features among the morphs *kleid* and *ung*. This correlation indicates that morphs can potentially be attributed to morphological features in an empirical sense, and by using such collocation measures it is possible to extract some informative signals.

Morph	Feature	ΔP forward	ΔP backward
ung	Case=Nom	0.129	0.226
ung	**Gender=Fem**	**0.467**	**0.798**
ung	Number=Sing	0.267	0.549
ung	Case=Dat	0.230	0.389
ung	Case=Acc	0.246	0.364
ung	Gender=Masc	-0.175	-0.230
ung	Case=Gen	0.152	0.116

Table 2: Probabilities of the morph *ung* in German.

In the case of Hungarian (Table 3), the morph *ek* has the strongest affinity for the feature Number=Plur. But there are other morphs too in Hungarian which are responsible for carrying the feature Number=Plur, like *ok*, *ak*, *ei* and *ai*. In the case of German too, there are multiple morphs (Table 4) that mark for the feminine gender, like *keit*, *schaft*, *enz*, and so on. Our current unsupervised approach successfully captures all the morphs attributed to a given morphological feature; we however reiterate that this finding is purely empirical given the available data resource.

Morph	Feature	ΔP forward	ΔP backward
ek	Case=Nom	-0.006	-0.146
ek	Number=Sing	-0.033	-0.431
ek	Person=3	0.031	0.427
ek	Definite=Ind	0.026	0.328
ek	PronType=Ind	0.064	0.099
ek	Mood=Ind	0.030	0.340
ek	Tense=Pres	0.032	0.344
ek	VerbForm=Fin	0.028	0.333
ek	Voice=Act	0.028	0.333
ek	**Number=Plur**	**0.163**	**0.531**

Table 3: Probabilities of the morph *ek* in Hungarian.

Morph	f(morph,feature)	ΔP forward	ΔP backward
ion	2	0.001	0.686
keit	59	0.053	0.697
heit	38	0.034	0.693
schaft	58	0.052	0.697
ung	**497**	**0.467**	**0.798**
enz	1	0.001	0.685

Table 4: Morphs for Gender=Fem in German.

From Table 5 and Table 6, we infer that the morphs *tunk* and *ok* both encode the features Number=Plur and Person=1 in Hungarian. In the case of verbs conjugated in first person plural like *voltunk* 'we were' and *tanultunk* 'we studied' the morph *tunk* has the feature set Number=Plur|Person=1, whereas the morph *ok* has the feature Number=Plur for nouns and Number=Plur|Person=2 for verbs (as in *tanultatok* 'you all studied'), as well as the feature Person=1 (e.g. in *tanulok* 'I study').

Morph	f(morph,feature)	ΔP forward	ΔP backward
tunk	**1**	**0.033**	**0.972**
ok	**7**	**0.232**	**0.852**
ak	5	0.165	0.690
ek	5	0.163	0.531
ai	1	0.033	0.972

Table 5: Morphs for Number=Plur in Hungarian.

We do observe that a morph in Hungarian or any other language may take on multiple grammatical functions; we only cite these selected examples to highlight how polysemous morphs can be. Based on these feature sets extracted from UD it is possible to explore all the grammatical functions handled by the morphs across languages.

Based on the ΔP scores, we find that the morphological features more strongly associated with the Latin morph *us* are Case=Nom, Gender=Masc and Number=Sing (Table 7). The other features

Morph	f(morph,feature)	ΔP forward	ΔP backward
tunk	**1**	**0.143**	**0.994**
ok	**1**	**0.136**	**0.119**
om	1	0.141	0.328
tam	1	0.143	0.994
ttem	1	0.143	0.994

Table 6: Morphs for `Person=1` in Hungarian.

Morph	Feature	ΔP forward	ΔP backward
us	**Case=Nom**	**0.018**	**0.349**
us	Case=Acc	-0.012	-0.247
us	Case=Dat	-0.013	-0.130
us	Degree=Cmp	-0.012	-0.044
us	**Gender=Masc**	**0.017**	**0.369**
us	Gender=Fem	-0.018	-0.346
us	Gender=Neut	-0.013	-0.262
us	**Number=Sing**	**0.014**	**0.159**
us	Number=Plur	-0.018	-0.349

Table 7: Probabilities of the morph *us* in Latin.

attributed to the morph *us* are potentially due to the feature values of the lexical root morph it happens to co-occur with. The ΔP backward scores indicate the morph *us* has a strong correspondence with the feature `Gender=Masc`.[17]

Given the observations, ΔP proves to be a strong unsupervised measure that extracts features associated with morphs, which potentially indicates that morphs do carry morphological features and in any case it would be reasonable to use this information to analyze word-internal structure in more detail.

4.3. Conforming to UD

When morphologically segmenting the nodes of a treebank, a natural question that arises concerns how to annotate morphs within UD. Specifically, when creating the morphological MWT we need to assign to its elements lemma, POS, morphological features, and deprel.

In many cases when segmentation is provided, UniSegments also comprises information about morphemes; namely, a word morph is possibly associated with its corresponding morpheme. For instance, the Latin verb *auerto* 'to turn away' is split as *a* + *uerto*, with the morph *a* associated to the morpheme *a(b)*, which can indeed take both forms *a* and *ab*. When available, we adopt the provided morpheme as a lemma; otherwise, we set the morph lemma to be identical to its form. We assign the POS that the node originally has (i.e., before undergoing the segmentation) to the head of MWT, which should correspond to the stem of the word. All other tokens of the MWT, i.e. morphs, receive the POS tag X. Indeed, we decide not to tag them with labels describing their position with respect to the stem (e.g., prefix, suffix) or the morphological process they convey (e.g., inflection, derivation). By assigning the X UPOS tag, we try to be as compliant as possible to UD, although without affirming that we believe morphs to have a POS.

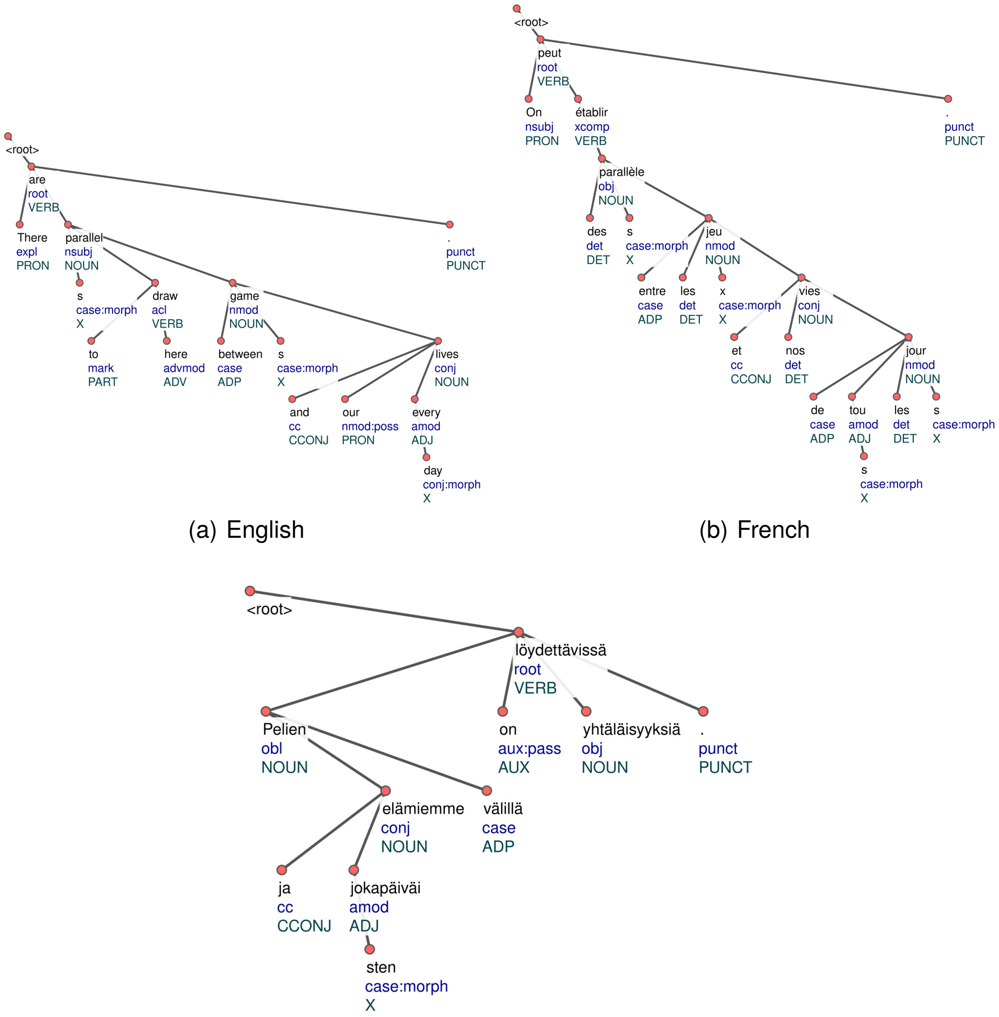
To annotate features, we exploit the feature-based alignment presented in Subsection 4.2. Specifically, for each of the morphological segments that we identify, we search for the features that are associated with them as a result of the alignment process. If any of those features can also be found in the original feature set of the token, we assign it to the morph and remove it from the set of features of the root, as we believe it to belong to the morph instead of the root.

When assigning deprels, we handle prefixes, root(s), and suffixes in a slightly different manner. Prefixes, extracted from UniSegments, are assigned `nmod:morph` if they are substantives (NOUN/PROPN), `advmod:morph` for all other POSs. If according to UniSegments the lemma presents just a single root, it inherits the deprel that the node originally had. If more than one is found, the second (and possibly more) is annotated as `conj:morph`. It is the case of compounds, for which the choice of `conj` is justified by the fact that we want all the lexical stems to be somehow on the same level. We are aware that parataxis is not the only possible relation between words constituting a compound (cf. Svoboda and Ševčíková 2024); however, we adopt this practical solution since the type of compound structure is not annotated in the exploited resources. As of now, we intend to use `conj:morph` only as a way to point out the co-existence of two lexical roots. In the case of suffixes, we try to approximately distinguish verbal and nominal inflection. Segmented morphs of verbs and auxiliaries are assigned `aux:morph`, while `case:morph` applies to nouns, adjectives, determiners, pronouns, adverbs, numerals, and extremely rare instances of adpositions. Whenever we are not able to reasonably assign either of the two deprels, we opt for `dep:morph`. As mentioned in the previous subsection, the `:morph` subtype allows to distinguish and retrieve all instances of morphological segmentations.

5. MorphoTrees

Figures 4(a), 4(b), and 4(c) display the same sentence, corresponding to English *There are parallels to draw here between games and our everyday lives*. The sentence, extracted from PUD treebanks, is shown also in Finnish and French and provides an example of how including the internal structure of

[17] However, this correlation comes purely from the data we have in hand. Theoretically, the morph *us* in Latin can equally express e.g. `Case=Nom`, `Gender=Masc`, and `Number=Sing`. Currently, we do not have a baseline to compare our empirical findings with theoretical facts.



(a) English

(b) French

(c) Finnish

Figure 4: UD-morphological tree of the sentence *There are parallels to draw here between games and our everyday lives* in English, French, and Finnish respectively.

words into UD could provide interesting remarks. Indeed, parallel data available in PUD could be observed in an even more parallel perspective after morph splitting, as in different languages some features could be realized differently, but a similar approach could help align them. In Appendix A we also display the raw CoNLL-U representation of the sentences (Figures 7, 8, 9), in order for the features and the MWT-like strategy to be visible.

In the Finnish example in Figure 4(c), the word form *jokapäiväisten* 'everyday ones' is split as *jokapäiväi* and *sten*. *Jokapäiväi* gets the POS tag ADJ and the deprel `amod` and the morph *sten* gets the deprel `case:morph` as decided. In the English example in Figure 4(a), the word form *games* is split as *game* and *s* where the morph *s* gets the deprel `case:morph`. The compound *everyday* is split and *day* is attached as `conj:morph` to *every*.[18] Similar splits can be also observed in the French example in Figure 4(b).[19] Figures 5 and 6 show the integration of segmentation within non-PUD treebanks.

[18] *Everyday* clearly shows a case where the two elements of the compound are attached paratactically according to our solution, whereas *every* is actually dependent on *day* within the structure of the compound.

[19] The example can also serve to highlight how the segmentation of the exploited resources, and hence its quality and level of granularity, is inherited in our data. For instance, in the verb *établir* the infinitive marker *ir* should be segmented, while it is not. Of course, this kind of choice also strongly depends on the adopted approach to morphological segmentation, which is far from being a solved problem yet. A similar observation would probably apply to Finnish as well, where some expected segmentations may be missing.

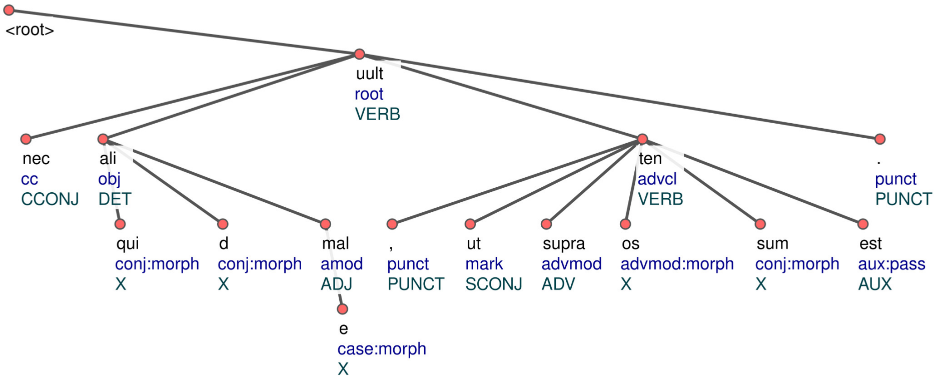


Figure 5: UD-morphological tree of the Latin sentence *Nec aliquid male uult, ut supra ostensum est.* ('Nor does he will anything evil, as we have proved.').

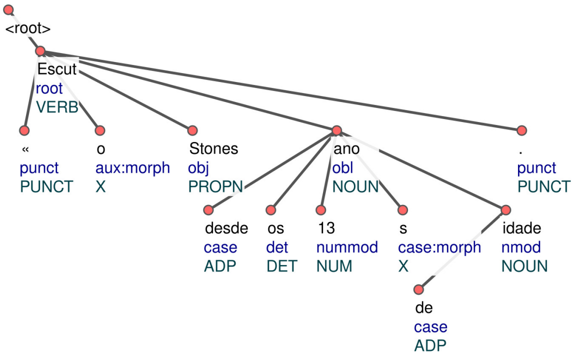


Figure 6: UD-morphological tree of the Portuguese sentence *Escuto Stones desde os 13 anos de idade.* ('I've been listening to the Stones since I was 13.').

6. Conclusion and Future Work

In the paper, we presented the proposal of a novel data structure aiming at integrating the representation of the morphological internal structure of words into Universal Dependencies. Working on 10 languages as a case study, we first devised a prototype of a methodology to manipulate UD treebanks intending to include the morphological structure of words into the canonical UD-like sentence representation. Then, we investigated the alignment between morphs and feature sets, by calculating ΔP scores that indicate the strength of the relation between a morph and a feature, and proceeded to assign relevant morphological features to morphs. Both tasks exploited already existing resources to perform segmentation. Such an approach ties the quality of our data to that of the resources we employed, for which some limitations were observed (derived e.g. from conversion from different resources).

Overall, the work we presented does not intend to suggest a reorganization of Universal Dependencies towards the inclusion of internal, morphological word structure. Our goal is to provide a resource that integrates morphology and syntax, two linguistic layers often intertwining, and that can prove beneficial in enhancing comparability of languages that express comparable meaning through different grammatical strategies[20]. The key factor for enhancing comparability lies in the cross-lingual correspondence of morphs.

In the future, we plan to improve the described workflow and expand the collection of morphological treebanks to more languages. Additionally, the extraction of the morphological trees from the sentence representation could be explored, towards their possible integration with DeriNet (Vidra et al., 2021). Moreover, in recent developments, morphological features are used to create multilingual morphological analyzers, for instance as presented by Pawar et al. (2023). We would like to carry forward our current research in that direction too by including a larger set of languages, as well as by including phenomena that we have neglected so far, such as non-concatenative morphology. We will find ways to estimate the quality of the resulting trees.

7. Acknowledgements

This work has been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062). The study was supported by the Charles University, project GA UK No. 104924 and project GA UK No. 101924; and partially supported by SVV project number 260 698. We would like to thank three anonymous reviewers for their very insightful feedback.

[20] Most notably, different degrees of inflection.

8. Bibliographical References

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria. Association for Computational Linguistics.

Barli Bram. 2012. Three models of English morphology. *LLT Journal: A Journal on Language and Language Teaching*, 15(1):179–185.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Robert M. Fano and David Hawkins. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies [Conversion and Improvement of Universal Dependencies French corpora]. *Traitement automatique des langues*, 60(2):71–95.

Jan Hajič, Eduard Bejček, Jaroslava Hlavacova, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Martin Haspelmath. 2020. The morph as a minimal linguistic form. *Morphology*, 30(2):117–134.

Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. In *Linguistic Issues in Language Technology, Volume 11, 2014-Theoretical and Computational Morphology: New Trends and Synergies*.

Herbert M. Jenkins and William C. Ward. 1965. Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, 79(1):1–17.

D. Jurafsky and J.H. Martin. 2014. *Speech and Language Processing*. Always learning. Pearson.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.

Zhongguo Li. 2011. Parsing the internal structure of words: A new paradigm for Chinese word segmentation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1414, Portland, Oregon, USA. Association for Computational Linguistics.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. Building a word formation lexicon for Latin. In *CLiC-it/EVALITA*.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David

Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Amir More, Özlem Çetinoğlu, Çağrı Çöltekin, Nizar Habash, Benoît Sagot, Djamé Seddah, Dima Taji, and Reut Tsarfaty. 2018. Conll-ul: Universal morphological lattices for universal dependency parsing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. *Digital Classical Philology*, 10:299–320.

Siddhesh Pawar, Pushpak Bhattacharyya, and Partha Talukdar. 2023. Evaluating cross lingual transfer for morphological analysis: a case study of Indian languages. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 14–26, Toronto, Canada. Association for Computational Linguistics.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal Dependencies for Finnish. In *Proceedings of NoDaLiDa 2015*, pages 163–172. NEALT.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy.

Claudia H Sánchez-Gutiérrez, Hugo Mailhot, S Hélène Deacon, and Maximiliano A Wilson. 2018. MorphoLex: A derivational morphological database for 70,000 English words. *Behavior research methods*, 50:1568–1580.

Ulrike Schneider. 2020. ΔP as a measure of collocation strength. Considerations based on analyses of hesitation placement in spontaneous speech. *Corpus Linguistics and Linguistic Theory*, 16(2):249–274.

Petra Steiner. 2017. Merging the trees - building a morphological treebank for German from two resources. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 146–160, Prague, Czech Republic.

Gregory T Stump. 2001. *Inflectional morphology: A theory of paradigm structure*, volume 93. Cambridge University Press.

Emil Svoboda and Magda Ševčíková. 2024. Compounds in Universal Dependencies: A survey in five European languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 88–99, St. Julian's, Malta. Association for Computational Linguistics.

Luigi Talamo, Chiara Celata, and Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure*, 9(1):72–102.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Veronika Vincze, Katalin Simkó, Zsolt Szántó, and Richárd Farkas. 2017. Universal Dependencies and morphology for Hungarian - and on the price of universality. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 356–365, Valencia, Spain. Association for Computational Linguistics.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, and Jonáš Vidra. 2022. Towards Universal Segmentations: UniSegments 1.0. In *Proceedings* *of the Thirteenth Language Resources and Evaluation Conference*, pages 1137–1149, Marseille, France. European Language Resources Association.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. Chinese parsing exploiting characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 125–134, Sofia, Bulgaria. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014. Character-level Chinese dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1326–1336, Baltimore, Maryland. Association for Computational Linguistics.

Hai Zhao. 2009. Character-level dependencies in Chinese: Usefulness and learning. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 879–887, Athens, Greece. Association for Computational Linguistics.

9. Language Resource References

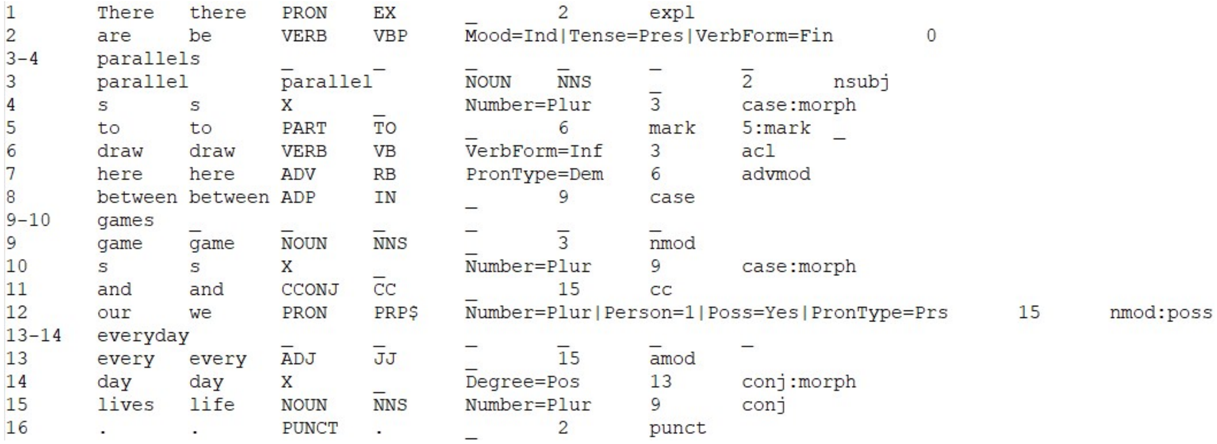
Baayen, R. Halard and Piepenbrock, Richard and Gulikers, Leon. 1995. *CELEX2*. Linguistic Data Consortium, ISLRN 204-698-863-053-1.

Vidra, Jonáš and Žabokrtský, Zdeněk and Kyjánek, Lukáš and Ševčíková, Magda and Dohnalová, Šárka and Svoboda, Emil and Bodnár, Jan. 2021. *DeriNet 2.1*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. PID http://hdl.handle.net/11234/1-3765.

Žabokrtský, Zdeněk and Bafna, Nyati and Bodnár, Jan and Kyjánek, Lukáš and Svoboda, Emil and Ševčíková, Magda and Vidra, Jonáš and Angle, Sachi and Ansari, Ebrahim and Arkhangelskiy, Timofey and Batsuren, Khuyagbaatar and Bella, Gábor and Bertinetto, Pier Marco and Bonami, Olivier and Celata, Chiara and Daniel, Michael and Fedorenko, Alexei and Filko, Matea and Giunchiglia, Fausto and Haghdoost, Hamid and Hathout, Nabil and Khomchenkova, Irina and Khurshudyan, Victoria and Levonian, Dmitri and Litta, Eleonora and Medvedeva, Maria and Muralikrishna, S. N. and Namer, Fiammetta and Nikravesh, Mahshid and Padó, Sebastian and Passarotti, Marco and Plungian, Vladimir and Polyakov, Alexey and Potapov, Mihail and Pruthwik, Mishra and Rao B, Ashwath and Rubakov, Sergei and Samar, Husain and Sharma, Dipti Misra and Šnajder, Jan and Šojat, Krešimir and Štefanec, Vanja and Talamo, Luigi and Tribout, Delphine and Vodolazsky, Daniil and Vydrin, Arseniy and Zakirova, Aigul and Zeller, Britta. 2022. *Universal Segmentations 1.0 (UniSegments 1.0)*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

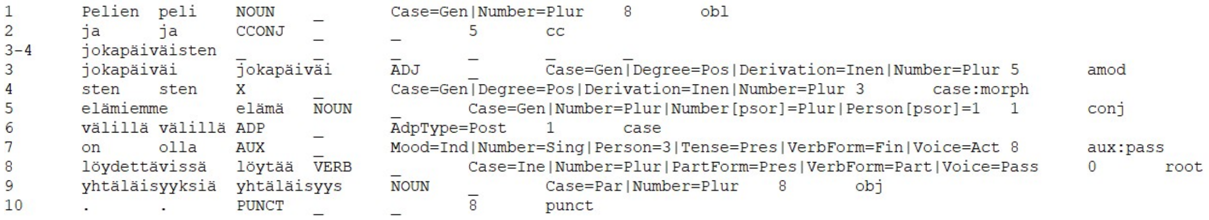
Zeman, Daniel et al. 2023. *Universal Dependencies 2.12*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. PID http://hdl.handle.net/11234/1-5150.

A. Example sentences in ConLL-U format



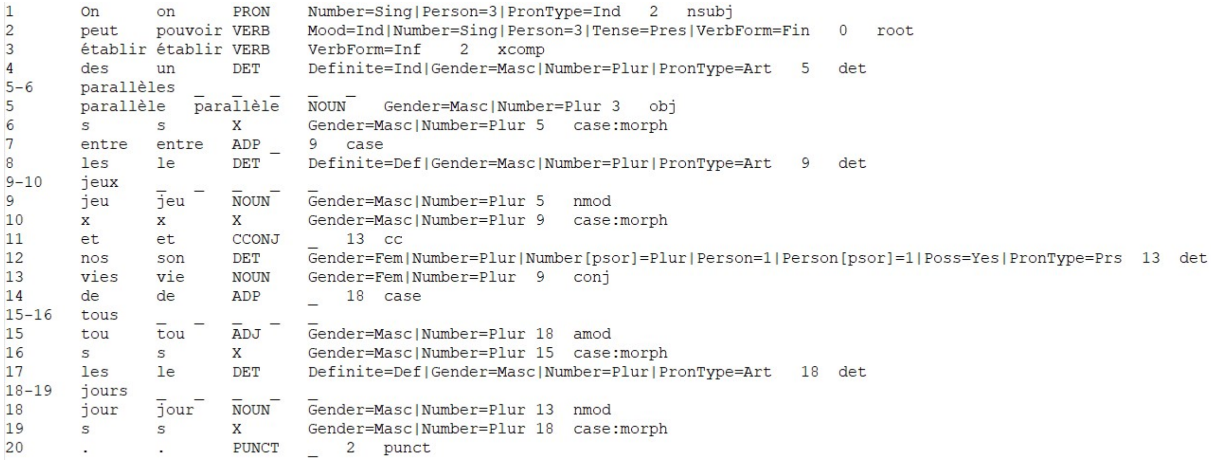
```
1       There   there   PRON    EX      _       2       expl
2       are     be      VERB    VBP     Mood=Ind|Tense=Pres|VerbForm=Fin        0
3-4     parallels       _       _       _       _       _       _
3       parallel        parallel        NOUN    NNS     _       2       nsubj
4       s       s       X       _       Number=Plur     3       case:morph
5       to      to      PART    TO      _       6       mark    5:mark  _
6       draw    draw    VERB    VB      VerbForm=Inf    3       acl
7       here    here    ADV     RB      PronType=Dem    6       advmod
8       between between ADP     IN      _       9       case
9-10    games   _       _       _       _       _       _
9       game    game    NOUN    NNS     _       3       nmod
10      s       s       X       _       Number=Plur     9       case:morph
11      and     and     CCONJ   CC      _       15      cc
12      our     we      PRON    PRP$    Number=Plur|Person=1|Poss=Yes|PronType=Prs      15      nmod:poss
13-14   everyday        _       _       _       _       _       _
13      every   every   ADJ     JJ      _       15      amod
14      day     day     X       _       Degree=Pos      13      conj:morph
15      lives   life    NOUN    NNS     Number=Plur     9       conj
16      .       .       PUNCT   .       _       2       punct
```

Figure 7: CoNLL-U representation of the English sentence *There are parallels to draw here between games and our everyday lives* (see also 4(a), 4(c), 4(b)). All three figures in the appendix allow us to better understand how morphological features have been treated. In the CoNLL-U files shown here the ninth and tenth fields have been removed, for reasons of space, as they are not strictly relevant to what is discussed in the present work.



```
1       Pelien  peli    NOUN    _       Case=Gen|Number=Plur    8       obl
2       ja      ja      CCONJ   _       _       5       cc
3-4     jokapäiväisten  _       _       _       _       _       _
3       jokapäivä       jokapäivä       ADJ     _       Case=Gen|Degree=Pos|Derivation=Inen|Number=Plur 5       amod
4       sten    sten    X       _       Case=Gen|Degree=Pos|Derivation=Inen|Number=Plur 3       case:morph
5       elämiemme       elämä   NOUN    _       Case=Gen|Number=Plur|Number[psor]=Plur|Person[psor]=1   1       conj
6       välillä välillä ADP     _       AdpType=Post    1       case
7       on      olla    AUX     _       Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 8       aux:pass
8       löydettävissä   löytää  VERB    _       Case=Ine|Number=Plur|PartForm=Pres|VerbForm=Part|Voice=Pass     0       root
9       yhtäläisyyksiä  yhtäläisyys     NOUN    _       Case=Par|Number=Plur    8       obj
10      .       .       PUNCT   _       _       8       punct
```

Figure 8: CoNLL-U representation of the Finnish sentence.



```
1       On      on      PRON    Number=Sing|Person=3|PronType=Ind   2   nsubj
2       peut    pouvoir VERB    Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root
3       établir établir VERB    VerbForm=Inf    2   xcomp
4       des     un      DET     Definite=Ind|Gender=Masc|Number=Plur|PronType=Art   5   det
5-6     parallèles _ _ _ _ _
5       parallèle   parallèle   NOUN    Gender=Masc|Number=Plur 3   obj
6       s       s       X       Gender=Masc|Number=Plur 5   case:morph
7       entre   entre   ADP _   9   case
8       les     le      DET     Definite=Def|Gender=Masc|Number=Plur|PronType=Art   9   det
9-10    jeux    _ _ _ _ _
9       jeu     jeu     NOUN    Gender=Masc|Number=Plur 5   nmod
10      x       x       X       Gender=Masc|Number=Plur 9   case:morph
11      et      et      CCONJ   _   13  cc
12      nos     son     DET     Gender=Fem|Number=Plur|Number[psor]=Plur|Person=1|Person[psor]=1|Poss=Yes|PronType=Prs  13  det
13      vies    vie     NOUN    Gender=Fem|Number=Plur  9   conj
14      de      de      ADP     _   18  case
15-16   tous    _ _ _ _ _
15      tou     tou     ADJ     Gender=Masc|Number=Plur 18  amod
16      s       s       X       Gender=Masc|Number=Plur 15  case:morph
17      les     le      DET     Definite=Def|Gender=Masc|Number=Plur|PronType=Art   18  det
18-19   jours   _ _ _ _ _
18      jour    jour    NOUN    Gender=Masc|Number=Plur 13  nmod
19      s       s       X       Gender=Masc|Number=Plur 18  case:morph
20      .       .       PUNCT   _   2   punct
```

Figure 9: CoNLL-U representation of the French sentence.

Combining Grammatical and Relational Approaches. A Hybrid Method for the Identification of Candidate Collocations from Corpora

Damiano Perri*, **Irene Fioravanti**†, **Osvaldo Gervasi***, **Stefania Spina**†
*University of Perugia
†University for Foreigners of Perugia
damiano.perri, osvaldo.gervasi@unipg.it
irene.fioravanti, stefania.spina@unistrapg.it

Abstract

We present an evaluation of three different methods for the automatic identification of candidate collocations in corpora, part of a research project focused on the development of a learner dictionary of Italian collocations. We compare the commonly used POS-based method and the syntactic dependency-based method with a hybrid method integrating both approaches. We conduct a statistical analysis on a sample corpus of written and spoken texts of different registers. Results show that the hybrid method can correctly detect more candidate collocations against a human annotated benchmark. The scores are particularly high in adjectival modifier relations. A hybrid approach to candidate collocation identification seems to lead to an improvement in the quality of results.

Keywords: Collocation, Automatic identification, Learner dictionary

1. Introduction

Multi-word expressions (henceforth, MWEs), defined as lexical units (collocations, idioms, lexical bundles, etc.) consisting of two or more words, have been the focus of extensive research in many areas including lexicography and NLP for several decades (Evert, 2004; Paquot, 2015; Spina, 2020). The creation of lexicographical combinatory resources, such as dictionaries of collocations, explicitly targeted to learners of second languages (L2s), has been undertaken mainly for English (McIntosh et al., 2002); (Rundell, 2010), although general dictionaries of collocations not explicitly addressed to L2 learners exist for several languages, including English (Benson et al., 1986), and Italian (Urzì, 2009; Tiberii, 2012; Lo Cascio, 2013). The use of language corpora has significantly boosted research on MWEs and their lexicographic applications. This is particularly evident in the area of lexicography dedicated to MWEs, where the identification of typical word combinations hugely benefits from the use of vast collections of texts. These corpora allow to extract frequent naturally occurring lexical patterns, with the aid of NLP and statistical techniques for the analysis of word combinations (Hanks, 2012).

Two main tasks are involved in the extraction of MWEs from corpora (Seretan, 2011): the automatic identification of candidates, often according to specific a priori criteria on their grammatical and/or syntactic patterns, and the detection of phraseologically meaningful combinations (collocations, in this case), often based on frequency and/or statistical association measures, to filter out sequences of words without phraseological relevance. In this study, our focus is on the first task of automatically identifying candidate collocations in Italian corpora. We assume that the effectiveness of the subsequent stages in creating a learner dictionary of collocation strongly depends on how accurate this candidate identification proves to be. The more an automatic system based on NLP techniques can accurately identify word combinations that are potential collocations, the more accurate the data on their frequency. As a consequence, the association measures used to filter out non-collocations, all of which are, to varying degrees, dependent on frequency, can benefit from more reliable frequency values, resulting in increased accuracy.

The present study reports on an experiment aimed at proposing a hybrid approach to this task by comparing and evaluating the two most commonly used candidate detection methods - the POS-based method and the syntactic dependency-based method - with a third one resulting from the integration of the two previous approaches. For the first two methods, we adopt the denomination from (Castagnoli et al., 2016) and refer to the POS-based as the P-based approach and the dependency-based as the S-based approach, while we refer to the third integrated method as the Hybrid approach. Current collocation extraction approaches rely on linguistic pre-processing (e.g., POS-tagging or dependency parsing) of source corpora to better identify the candidates (Seretan, 2011). Previous research has shown that the P-based and S-based approaches have some limitations. The former re-

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@LREC-COLING-2024, pages 138–146. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

lies on an accurate and established NLP task such as POS-tagging. However, relying on positional POS patterns, it fails to capture the syntactic relations between word pairs or the marked sentence structures where the regular constituent order is reversed. For instance, a P-based approach would not detect the verb-direct object relation between *play* and *role* in Example 1 (the example is taken from Seretan, 2011, 59).

Example 1. *It is true, we must combat the menace of alcoholism in young people, and this text successfully highlights the* ***role*** *that families, teachers, producers and retailers must* ***play*** *in this area.*

On the contrary, this relation would probably be detected using an S-based approach that relies on parsed data and thus can identify the verb-direct object dependency. Another advantage of this approach is that it does not limit the distance between the two words constituting the candidate collocation, unlike the P-approach. However, parsing errors are a well-known shortcoming of this approach: error rates ranging from 7.85% to 9.7% of the total candidate collocations extracted were reported to be due to parsing errors by previous studies (Wu and Zhou, 2003; Lin, 1999). Despite the recent improvement in parsing accuracy, (Qi et al., 2020; Akbik et al., 2018) the parsing approach still has limitations in selecting candidate collocations as it provides little information on how words combine with each other and fail to distinguish frequent combinations and idiomatic ones with the same syntactic structure (Castagnoli et al., 2016).

This study aims to present a hybrid approach to detecting candidate collocation from corpora for lexicographic applications on a language different from English, i.e. Italian. The hypothesis we aim to validate is that this hybrid approach performs better in the candidate identification task. From an exploratory perspective, we also intend to investigate cases in which the hybrid method works better and identify cases where further improvements might be warranted.

2. Related work

In this section, we briefly survey the main methods and NLP techniques used to perform the specific task of detecting, or discovering (Constant et al., 2017) candidate collocations from corpora, regardless of the measures employed to identify the proper phraseological collocations, which represents a further step in the process of assembling the set of entries required by the lexicographic application.

Early NLP works addressing this task identified candidate collocations using frequent word sequences, regardless of their syntactic structure, and relied on n-gram methods to extract them from corpora (Choueka, 1988; Smadja, 1993). Later, this search "for needles in a haystack" (Choueka, 1988) more and more employed linguistically preprocessed corpora and lemmatised and POS-tagged data. This further step was especially suitable for handling morphological and syntactic variability typical of languages with richer morphology and more accessible word order (Evert, 2004). The P-approach is the first to become established, given the widespread availability of POS-tagged corpora in many languages. Many extraction systems relying on this approach involve an a priori selection of specific types of POS combinations (e.g. verb-noun, adjective-noun, etc.). Right from the start, a drastic improvement in the detection accuracy was found when a POS filter was applied (Breidt, 1993; Daille, 1994; Krenn, 2000; Ritz, 2006). These results were primarily reported for fixed and adjacent candidates, where even a simple linguistic analysis can capture basic grammatical patterns.

In later years, it has been suggested that the detection of candidate collocations can benefit from a finer linguistic analysis of texts. Seretan's (2011) extensive study explored and evaluated the use of syntactic dependencies, as they can also capture discontinuous and syntactically flexible candidate collocations based on syntactic relations between words, improving the quality of the results. However, many systems relying on an S-approach aimed at MWE identification after parsing, so as to benefit from the previous syntactic analysis (Constant et al., 2017) reported high parsing error rates affecting the accuracy of the detection task. The issue of parsing accuracy is identified and evaluated by several studies (e.g. Orliac and Dillinger, 2003; Lü and Zhou, 2004). Lü and Zhou (2004) identified a parsing error rate >7%. Orliac and Dillinger (2003) also evaluated the most recurrent parsing errors and found that relative constructions were responsible for nearly half of the candidate collocations missed by their system.

Given all these reported limitations, it can be argued that the existing detection methods relying on an S-based approach are promising but have not yet been fully developed, due to issues related to parsing accuracy. There is, therefore, a general call for hybrid approaches to candidate collocation detection, combining the advantages of both P-based and S-based approaches while minimising their shortcomings. As Castagnoli et al. (2016) claimed, "the two methods seem to be highly complementary rather than competing with one another". Some attempts have been made to integrate the two approaches in recent years. Simkó et al. (2017) proposed a system using both POS-tagging and dependency parsing to identify single- and multi-token verbal MWEs in texts and reported the best results on the verb-particle constructions where their system correctly identified around 60% of constructions, but only about 40% of other types. Shi and Lee (2020) proposed a joint method that combines scores from both POS-tagging and dependency parsing to extract headless MWEs. Their results showed that tagging is more accurate than parsing for identifying flat-structure MWEs. At the same time, the joint method leads to higher accuracy, and most of the gains derive from shared results between parsers and taggers.

3. Method

To validate our hypothesis and explore the performance of different systems in automatically detecting candidate collocations in Italian corpora, we designed our experiment to mimic the "natural" processes that will be employed in the final extraction of candidates to be included in a learner dictionary of Italian collocations. For instance, we did not preselect target words or lemmas for the experiment. Instead, we considered all the word pairs produced in a text sample.

The only pre-selection we made was the syntactic relations of the candidate collocations. We opted to focus on syntactically-bound combinations, as the task of detecting candidate collocations is targeted to a lexicographic application. In the final dictionary entries, these collocations will be presented in accordance with their syntactic patterns. The choice was to investigate the two dependencies verb + direct object (Vdobj) and adjective modifier (amod) before and after a noun (both word orders are allowed in Italian). The choice is motivated by reasons of coverage and diversification. Firstly, previous research has shown that, among the eight syntactic structures most commonly forming collocations in Italian (verb + direct object, amod, noun + preposition + noun, noun + noun, verb + adjective, verb + adverb, noun + conjuction + noun, adjective + conjunction + adjective), the two that are considered in this study (Vdobj and amod) cover more than 50% of the total structures (Spina, 2016). Furthermore, while in both relations the order of the two components can be reversed, they have different features in terms of distance between their two components. In the Vdobj word combinations the distance between the two components can be even of several words (Example 2: there are five words between the verb *mantenere* 'keep' and the direct object *promesse* 'promises', and the two words are connected by a relative pronoun), while in the case of amod the two words are usually adjacent (Example 3) or near adjacent (Example 4).

Example 2. Non fare **promesse** che non riuscirai mai a **mantenere**!
Don't make promises you will never keep!

Example 3. Elisa mi stava raccontando della sua **brutta avventura**
Elisa was telling me about her bad adventure

Example 4. Questo è il **momento** più **atteso** della giornata
This is the most awaited moment of the day

3.1. Sample texts

We randomly extracted eight texts from a reference corpus of Italian, the *Perugia corpus* (Spina, 2014; https://lt.eurac.edu/cqpweb/), of the total size of ca. 8,000 tokens, balanced across written (tokens = 4,000) and spoken (tokens = 4,000) registers. We included different text genres: two newspaper articles (a report and an editorial), two school essays and a tourism-related blog post for the written part, and transcriptions of a conference, of a political speech and of the dialogues of a television series for the spoken part. On the one hand, this diversification in registers and text genres allows us to perform a simulation close to the actual extraction of candidate collocations for all the combination types in the whole corpus. On the other hand, it enables us to evaluate the three approaches to this task for register variation, which could affect accuracy.

3.2. The three systems

We used the systems described below to compare three different methods for detecting and extracting candidate collocations from Italian corpora, whose output was compared with a benchmark of human annotation.

P-based approach The sample texts were POS-tagged using *TreeTagger* (Schmid, 1994), trained with an ad hoc tagset based on a fine-grained set of 54 POS tags (Spina, 2014). Afterwards, the texts were searched via the *Corpus Workbench* (CWB) tool (Hardie, 2012) and the *Corpus Query Processing* (CQP) system by using three separate queries to detect the Vdobj relations and the two positional variants of the amod relations, with the adjective preceding or following the modified nouns. The three queries integrate POS tag sequences (the target ADJ, NOUN and VERB POS tags, as well as those that can potentially be inserted within the two constituents of the combinations, like articles, conjunctions or adverbs) and regex with lemmas to exclude (a list of the most frequent intransitive Italian verbs). The direct output of this regex-over-pos process represents the P-based approach, that was able to identify 549 candidate collocations.

S-based approach In this approach, a candidate collocation consist of two syntatically related lexical items. Therefore, the main criterion for detecting a candidate is the presence of a syntactic relation between the two items, in our case, the Vdobj and amod relations. In addition, to be identified as a valid candidate, each pair must satisfy more specific grammatical constraints. For instance, the words involved in the syntactic relations can only be nouns, adjectives or verbs. The sample texts were parsed using the framework of Universal Dependencies for treebank annotation (UD; de Marneffe et al., 2021) and the popular open-source library for advanced NLP in Python *spaCy*. Artificial intelligence is to date applied in many areas of science (Benedetti et al., 2020; Perri et al., 2022; Milani et al., 2021). The *spaCy* library is an example of the application of artificial intelligence to linguistic analysis. Since the simple parsing output does not yet represent the S-approach, the complete procedure details are described in section 3.3. The final number of candidate collocations identified by the S-based approach is 685.

Hybrid approach The hybrid approach results from merging the two previous approaches. It includes all the common candidates identified by both, as well as those only detected by the P-based approach and those only detected by the S-based approach. The Hybrid approach identified 748 candidate collocations.

3.3. Annotation

The output of the three systems was compared to a benchmark obtained by human evaluation. Two Italian trained linguists manually extracted all the Vdobj and amod combinations used in the eight sample texts. The two human annotators only adopted the criterion of the syntactic relations to extract the candidate collocations. Without calculating the inter-annotator agreement, any inter-annotator disagreements were resolved through negotiation until consensus was achieved for all forms. This annotation process resulted in a list of 610 candidate collocations, which served as a benchmark for the following steps.

3.4. Computational procedure

Three steps make up the computational process, allowing consistent and thorough data processing. The preliminary pre-processing of the texts was first carried out to enable homogeneous treatment of information. In the second step, the sentences were parsed using *spaCy*, and a set of rules was implemented to optimise the analysis. Finally, the results were statistically treated. Specifically, the results obtained through the S-approach were compared to those obtained through the P-approach and the Hybrid approach.

3.4.1. The pre-processing of the input texts

The first step involved pre-processing the texts to standardise the input data format and remove any irrelevant elements for analysis. This process included inserting capital letters at the beginning of each sentence and full stops at the end. We removed all whitespace due to typing errors (e.g. double whitespace) or whitespace after the end of a sentence in order to ensure that all input is as clean and error-free as possible. The sentences were then extracted and inserted into a data structure. Each sentence was assigned to a row within a spreadsheet (CSV file), constituting the database for the following stages of the analysis. Having one sentence per line is crucial, as it ensures an easily repeatable analysis and prevents overloading the *spaCy* parser, which can operate with a limited amount of RAM without requiring excessive resources.

3.4.2. The parsing of input phrases

The second phase of our work was devoted to sentence parsing using *spaCy* and the rules implemented in Python to recognize adjective modifier dependency (amod) and verb-direct object dependency (Vdobj).

The syntactic analyzer is a Python object obtained by importing the pre-trained *spaCy* library on the CPU-optimized Italian pipeline called `it_core_news_lg`[1]. The pre-training model occupies 541MB of written text (news and media). The pipeline provided by the model consists of `tok2vec`, `morphologizer`, `tagger`, `parser`, `lemmatizer`, `attribute_ruler`, `ner`. *spaCy* was trained with the UD Italian ISDT v2.8 (Italian Stanford Dependency Treebank; Attardi et al., 2015) There are various software libraries that can be used to perform the task of analysing the grammar of a sentence. We opted for *spaCy* since a version of its Italian language model was released very recently, on 1 Oct 2023[2].

Each sentence in our corpus was analyzed word by word. Given a word, *spaCy* provides a list of output objects: DepRel, `Form`, `Lemma`, `UPosTag`, `XPosTag`, `head.i`.

- `DepRel`: indicates the syntactic dependence relationship of the word to the main word in the sentence.
- `Form`: represents the word's surface form and how it appears in the text.

[1] https://spaCy.io/models/it#it_core_news_lg

[2] https://github.com/explosion/spacy-models/releases/tag/it_core_news_lg-3.7.0

Table 1: Comparison of the performance metrics of the three models across the entire dataset.

	Accuracy	Recall	Precision	F1 Score	Benchmark Match
P-based	0.70	0.79	0.87	0.83	78.90%
S-based	0.67	0.86	0.75	0.80	85.88%
Hybrid	0.67	0.90	0.73	0.80	90.20%

Table 2: Comparison of performance metrics of the three models concerning modifier adjectives.

	Accuracy	Recall	Precision	F1 Score	Benchmark Match
P-based	0.76	0.83	0.90	0.87	83.43%
S-based	0.68	0.88	0.75	0.81	88.25%
Hybrid	0.70	0.93	0.73	0.82	93.37%

Table 3: Comparison of performance metrics of the three models concerning verb-object combination.

	Accuracy	Recall	Precision	F1 Score	Benchmark Match
P-based	0.63	0.73	0.82	0.77	73.33%
S-based	0.66	0.83	0.76	0.79	82.96%
Hybrid	0.64	0.86	0.71	0.78	86.30%

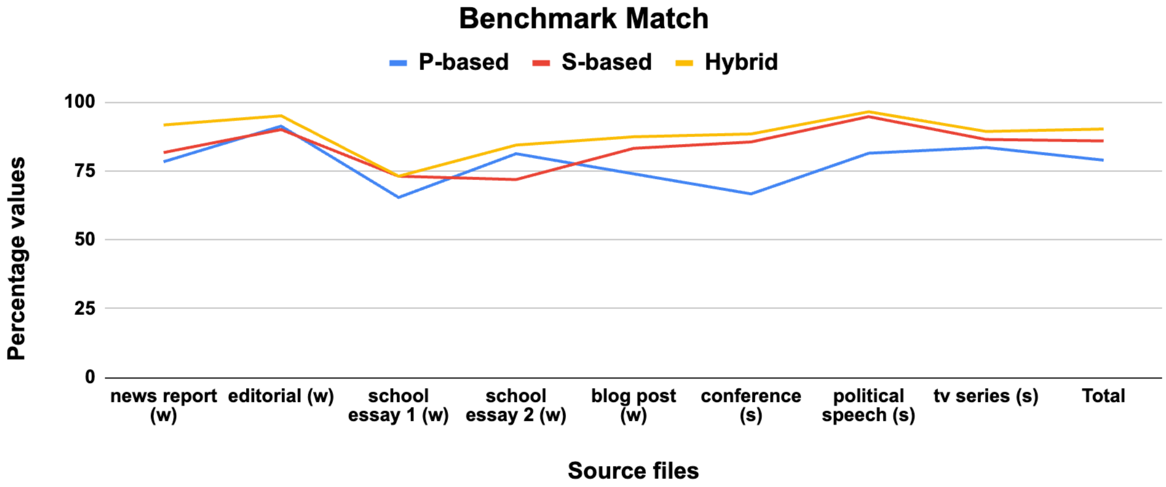


Figure 1: Benchmark Match values per file related to the entire dataset (w=written, s=speech).

- `Lemma`: is the basic form of a word that appears in dictionaries.

- `UPosTag` (Universal Part of Speech Tag): indicates the grammatical category of the word according to the universal POS tag scheme.

- `XPosTag` (Extended Part of Speech Tag): provides an extended POS tag that can include additional information.

- `head.i` (Head index): indicates the index of the word to which the current word is directly connected as a child in the sentence tree structure.

This information alone is not sufficient to fully understand the sentence's logical structure. Therefore, we identified several syntactic rules translated into Python functions to check the currently examined word and its head and determine whether it is part of an amode or Vdobj word combination. These rules were crucial in increasing the model's accuracy and precision, by cross-using the values of the different linguistic information provided by the parsing output. Writing these rules is particularly complex, as Italian is a morphologically and syntactically rich language with relatively free word order. For this reason, we proceeded step by step by analyzing the results obtained from time to time

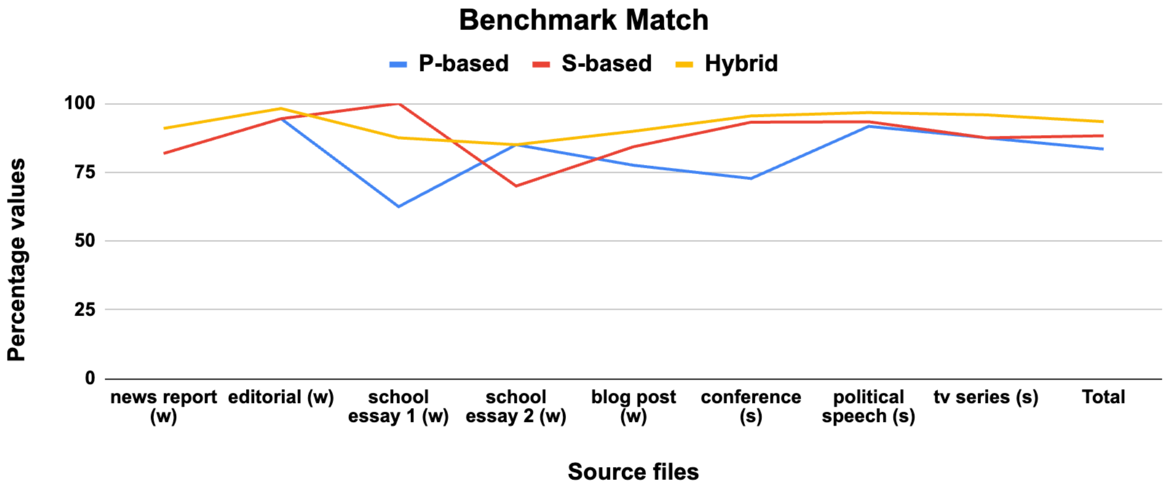


Figure 2: Benchmark Match values per file related to the modifier adjectives (w=written, s=spoken).

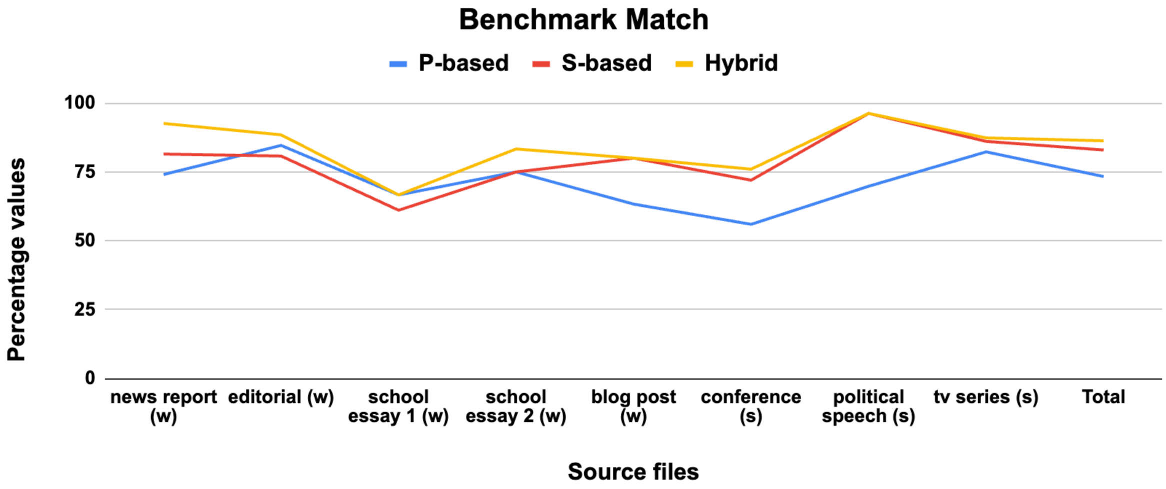


Figure 3: Benchmark Match values per file related to the verb-object combination (w=written, s=spoken).

and checking for incorrectly classified words to add rules, allowing the model to identify as many word combinations as possible. It is important to emphasize that the Python rules are specifically designed for the Italian language.

Some of the most important grammar rules that have been translated into Python code are now given. The first function recognizes a direct verbal object (Vdobj) with the obj relation with root as the dependency, while simultaneously verifying that the UPosTag of the root is VERB.

```
if token.dep_ == "obj" and
   token.head.dep_ == "ROOT"
   and token.pos_=="NOUN" and
   token.head.pos_ == "VERB"
```

This rule is able to recognize the combination of words *hanno fama* in Example 5.

Example 5. Molto note per le proprietà minerali delle acque sono le sorgenti di nitrodi e di olmitello, le loro virtù terapeutiche **hanno fama** mondiale. *Well-known for the mineral properties of the waters are the nitrodi and holmitello springs, their therapeutic virtues are world-renowned.*

Conversely, the function below is designed to identify AMOD when the 'amod' relation exists, with 'obj' as the dependency, and the UPosTag of the 'obj' token is NOUN.

```
if token.dep_ == "amod" and
   token.head.dep_ == "obj"
   and token.pos_=="ADJ" and
   token.head.pos_ == "NOUN"
```

The previous rule is able to recognize the word combination *straordinarie proprietà* in Example 6.

Example 6. Poi arrivarono i romani e scoprirono le **straordinarie proprietà** delle acque calde.
Then the Romans came and discovered the extraordinary properties of hot water.

In total, we created 18 functions to help us in identifying amod and Vdobj syntactic patterns. These functions were subsequently added to a function array. Each word was parsed from the function array, and upon finding a match, the result was saved in our data structure.

```
for token in line:
  for fun in functionsList:
    if fun(token):
      found="*"
```

At the end of this step, we obtained a data structure without duplicates of all word combinations categorized as amod or Vdobj, which was used as the input for the next step.

3.4.3. Statistical analysis of the model

The performance of the three approaches (P-based, S-based and Hybrid) was compared and evaluated through the usual measures of accuracy, precision, recall, F1 score. We defined in addition the *benchmark match*, which represents the percentage between the predictions generated by the model and the corresponding class labels in the benchmark file. It indicates how well the model aligns with the correct predictions established by the benchmark file, demonstrating its reliability and consistency against a validation dataset. The formula is $bm = 100 * (TP + TN)/(TP + TN + FN)$, where TP=True Positive, TN=True Negative, and FN=False Negative.

The Hybrid approach outperforms the P- and S-based approaches for the benchmark match and for recall. This better performance is observable across the entire dataset (Table 1), as well as for each of the syntactic relations taken individually (Tables 2 and 3). For the amod relation, the Hybrid approach reaches 93,37% of the benchmark match. This score can be regarded as highly positive in the context of candidate collocation identification. As expected, the P-based approach has better precision and worse recall, suggesting it has the lowest number of false positives but a reduced ability to identify positive instances. Conversely, the S-based approach shows low precision and high recall. It is worth noting that all the three methods have poorer results in detecting Vdobj relations compared to amod relations (Table 3), as in Vdobj relations the two words can be distant and in inverted order. However, the P-based approach is the one that has the most significant loss in benchmark match for Vdobj combinations (-10% compared to the amod relation).

In Figure 1, the benchmark match values related to the three approaches and the entire dataset are plotted as a function of the single sample files. Similar information is shown in Figure 2 about amod relation alone and in Figure 3 about Vdobj combinations alone. The figures allow for an evaluation of possible register influences on detection accuracy. The texts where the three approaches exhibit the most significant differences are two spoken texts, with a relatively formal register: the conference and the political speech, where the P-based approach has the worst results (Figure 1).

Overall, the Hybrid model validates our predictions and aligns more closely with the correct predictions established by the benchmark set, proving its reliability in complying with the gold standard of human annotation. The benefit of integrating the positional part of-speech and syntactic information for candidate collocation extraction is thus confirmed.

4. Conclusions and future work

Focusing on the automatic identification of candidate collocations in Italian corpora for lexicographic purposes, this study reports on an experiment aimed at comparing and evaluating the two most commonly used candidate detection approaches - the P-based and the S-based approach - with a third hybrid method resulting from the integration of the two previous ones. The evaluation of this step is crucial in order to assess the quality of candidate collocations with respect to specific criteria: their grammatical well-formedness (Seretan, 2011). Our assumption was that this quality would benefit from the integration of robust regex-over-pos methods with syntax-based approaches, despite the challenges posed by parsing large amounts of text in a morphosyntactically rich language like Italian. Results show that the Hybrid approach outperforms the two other methods in benchmark match and recall values, confirming the validity of our assumptions. Further work is still needed to optimise the model as precision, accuracy and F1 score obtain higher values with a P-based approach. By implementing additional Python rules, e.g. negative rules (i.e. rules capable of removing false positives) we believe we can enhance the performance of the S-based approach by refining the predictive accuracy while reducing false positives. This, when combined with the outcomes of the P-based approach, is expected to result in an overall enhancement in the model's performance.

Although the robustness of post-tagging can balance to some extent the lower accuracy of syntactic parsing, the rules applied in detecting syntactic relations after parsing need refinements to reduce errors resulting from false positives. One limitation of this experiment derives from using only two syntactic relations, whereas the final procedure for dictionary entry selection will need to consider a larger set of relations. However, the conclusion that can be drawn is that pursuing a hybrid approach to candidate collocation identification is worthwhile, as it leads to an improvement in the quality of results.

5. Acknowledgements

The research has been funded by the Italian Ministry of Research (MUR), PRIN: Research Projects of Major National Interest – Call 2022 - Prot. 2022HXZR5E. The title of the project is: *DICI-A: A Learner Dictionary of Italian Collocations*.

6. References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Giuseppe Attardi, Simone Saletti, and Maria Simi. 2015. Evolution of italian treebank and dependency parsing towards universal dependencies. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015 - Trento - 3-4 December 2015*, Torino. Accademia University Press.

Priscilla Benedetti, Damiano Perri, Marco Simonetti, Osvaldo Gervasi, Gianluca Reali, and Mauro Femminella. 2020. Skin cancer classification using inception network and transfer learning. *Lecture Notes in Computer Science*, 12249 LNCS:536 – 545. Green Open Access.

Morton Benson, Evelyn Benson, and Robert Ilson. 1986. *The BBI Dictionary of English Word Combinations*. Benjamins, Amsterdam.

Elisabeth Breidt. 1993. Extraction of v-n-collocations from text corpora: A feasibility study for german. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 74–83, Columbus, USA.

Sara Castagnoli, Gianluca E. Lebani, Alessandro Lenci, Francesca Masini, Malvina Nissim, and Lucia C. Passaro. 2016. Pos-patterns or syntax? comparing methods for extracting word combinations. In Gloria Corpas Pastor, editor, *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 116–128. Tradulex, Geneve.

Yaacov Choueka. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *User-Oriented Content-Based Text and Image Handling*, page 609–623, Paris, FRA. Le Centre de Hautes Etudes Internationales d’informatique Documentaire.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Béatrice Daille. 1994. *Approche mixte pour l’extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Ph.D. thesis, Université Paris 7.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Patrick Hanks. 2012. Corpus evidence and electronic lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*, pages 57—-82. Oxford University Press, Oxford.

Andrew Hardie. 2012. Cqpweb combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Brigitte Krenn. 2000. Collocation mining: Exploiting corpora for collocation idenfication and representation. In *Proceedings of KONVENS 2000*, Ilmenau, Germany.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317—-324, Morristown, NJ, USA.

Vincenzo Lo Cascio. 2013. *Dizionario Combinatorio Italiano*. Benjamins, Amsterdam.

Yajuan Lü and M. Zhou. 2004. Collocation translation acquisition using monolingual corpora. In *Annual Meeting of the Association for Computational Linguistics*.

Colin McIntosh, Ben Francis, and Richard Poole. 2002. *Oxford Collocations Dictionary for Students of English*. Oxford University Press, Oxford.

Alfredo Milani, Valentina Franzoni, and Giulio Biondi. 2021. Parsing tools for italian phraseological units. In *Computational Science and Its Applications – ICCSA 2021*, pages 427–435, Cham. Springer International Publishing.

Brigitte Orliac and Mike Dillinger. 2003. Collocation extraction for machine translation. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA.

Magali Paquot. 2015. Lexicographyand phraseology. In Douglas Biber and Randi Reppen, editors, *The Cambridge Handbook of English Corpus Linguistics*, Cambridge Handbooks in Language and Linguistics, pages 460–477. Cambridge University Press.

Damiano Perri, Marco Simonetti, and Osvaldo Gervasi. 2022. Synthetic data generation to speed-up the object recognition pipeline. *Electronics (Switzerland)*, 11(1). Gold Open Access.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Julia Ritz. 2006. Collocation extraction: Needs, feeds and results of an extraction system for german. In *Proceedings of the workshop on Multi-word-expressions in a multilingual context at the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 41–48, Trento, Italy.

Michael Rundell. 2010. *Macmillan Collocations Dictionary for learners of English*. Macmillan Education, London.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Violet Seretan. 2011. *Syntax-based collocation extraction*. Springer, Dordrecht.

Tianze Shi and Lillian Lee. 2020. Extracting headless mwes from dependency parse trees: Parsing, tagging, and joint modeling approaches. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8780–8794, Online. Association for Computational Linguistics.

Katalin Ilona Simkó, Viktória Kovács, and Veronika Vincze. 2017. Uszeged: Identifying verbal multi-word expressions with pos tagging and parsing techniques. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 48–53, Valencia, Spain. Association for Computational Linguistics.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143—-177.

Stefania Spina. 2014. Il perugia corpus: una risorsa di riferimento per l'italiano. composizione, annotazione e valutazione. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, volume 1, pages 354–359, Pisa. Pisa University Press.

Stefania Spina. 2016. Learner corpus research and phraseology in italian as a second language: The case of the dici-a, a learner dictionary of italian collocations. In Begoña Sanromán Vilas, editor, *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*, pages 219–244. Memoires de la Societe Neophilologique de Helsinki, Helsinky.

Stefania Spina. 2020. The role of learner corpus research in the study of l2 phraseology: main contributions and future directions. *Rivista di psicolinguistica applicata - Journal of Applied Psycholinguistics*, XX(2):35–52.

Paola Tiberii. 2012. *Dizionario delle collocazioni*. Zanichelli, Bologna.

Francesco Urzì. 2009. *Dizionario delle Combinazioni Lessicali*. Convivium, Luxemburg.

Hau Wu and Ming Zhou. 2003. Synonymous collocation extraction using translation information. In *Proceeding of the Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 120—-127.

Multiword Expressions between the Corpus and the Lexicon: Universality, Idiosyncrasy and the Lexicon-Corpus Interface

Verginica Barbu Mititelu[1] **, Voula Giouli**[2]**, Kilian Evang**[3]**, Daniel Zeman**[4]**, Petya Osenova**[5]**, Carole Tiberius**[6]**, Simon Krek**[7]**, Stella Markantonatou**[8]**, Ivelina Stoyanova**[9]**, Ranka Stankovic**[10]**, Christian Chiarcos**[11]

[1]Romanian Academy Research Institute for Artificial Intelligence, vergi@racai.ro
[2]Insitute for Language and Speech Processing, Athena Research Centre, voula@athenarc.gr
[3] Heinrich Heine University Düsseldorf, evang@hhu.de
[4] ÚFAL MFF, Charles University, zeman@ufal.mff.cuni.cz
[5]Institute of Information and Communication Technologies, BAS, petya@bultreebank.org
[6]Leiden University, c.p.a.tiberius@hum.leidenuniv.nl
[7]Jožef Stefan Institute, simon.krek@ijs.si
[8]Insitute for Language and Speech Processing, Athena Research Centre, marks@athenarc.gr
[9]Institute for Bulgarian Language, BAS, iva@dcl.bas.bg
[10]University of Belgrade, ranka@rgf.rs
[11]University of Augsburg, christian.chiarcos@uni-a.de

Abstract

We present ongoing work towards defining a lexicon-corpus interface to serve as a benchmark in the representation of multiword expressions (of various types – nominal, verbal, etc.) in dedicated lexica and the linking of these entries to their corpus occurrences. The final aim is the harnessing of such resources for the automatic identification of multiword expressions in a text. The involvement of several natural languages aims at the universality of a solution not centered on a particular language, and also accommodating idiosyncrasies. Challenges in the lexicographic description of multiword expressions are discussed, the current status of lexica dedicated to this linguistic phenomenon is outlined, as well as the solution we envisage for creating an ecosystem of interlinked lexica and corpora containing and, respectively, annotated with multiword expressions.

Keywords: multiword expression lexicon, corpus, proof-of-concept lexicon encoding

1. Introduction

In the last decade, the PARSEME COST Action (Savary et al., 2015) created the prerequisites for annotating corpora with multiword expressions (MWEs), mainly verbal ones. Consistent guidelines[1] and an infrastructure for ensuring annotation consistency were developed, while the interaction among the members of the community was made possible by the COST Action and extended even beyond its duration. A corpus was created for 26 languages (Savary et al., 2023), in which verbal MWEs (VMWEs) were annotated according to the established guidelines. Meanwhile, a new COST Action, UniDive[2], is gathering the community again, simultaneously increasing in size and allowing for the development of guidelines for annotating MWEs of other parts of speech, and eventually for further annotation of corpora with the new MWE types, as well as for increasing the number of languages represented in the corpus so far. At the same time, UniDive builds on Universal Dependencies (UD) (de Marneffe et al., 2021), which posits standardized guidelines for tokenization, lemmatization and morphosyntactic annotation in treebanks of languages.

Despite the abundance of large bodies of annotated corpora and large language models, systems still fail to adequately identify MWEs and thus the need for lexica that are specifically designed to handle MWEs within the context of Natural Language Processing (NLP) (Savary et al., 2019b). Within UniDive, Working Group 2[3] seeks to take this further and to schematize the steps needed towards creating an ecosystem in which annotated corpora and MWE lexica are linked together, intra- and interlingually and are used to facilitate MWE identification in a way that universality and idiosyncrasy are taken into account.

In this paper, we report on original (ongoing) work towards designing this lexicon-corpus interface. The paper is structured as follows: we first outline our goals and the challenges we need to face (Section 2); then, an overview of the current

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=030_Categories_of_VMWEs

[2] https://unidive.lisn.upsaclay.fr/

[3] https://unidive.lisn.upsaclay.fr/doku.php?id=wg2:wg2

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 147–153. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

MWE dedicated lexica and the results of a survey aimed at better accounting for universality are presented (Section 3 and Section 4 respectively). The initial steps towards designing the lexicon-corpus interface, in a standardized manner with all its advantages are presented in Section 5. We outline the minimal requirements for encoding MWEs in computational lexica, with an eye to their interlinking with annotated corpora, in Section 6. Our conclusion is presented in Section 7.

2. Towards a lexicon-corpus interface: goals and challenges

For many decades, MWE-aware lexica have contributed a much larger set of MWEs than (annotated) corpora can do, as MWEs are rather rare in texts (Savary et al., 2019a), and to model their linguistic properties, namely, non-compositionality, lexical fixedness, discontinuity, potential modifiers of components, word order variation, etc. However, the representation of MWEs in hand-crafted lexica is far from homogeneous and even incomplete. At the same time, annotated corpora have been used as major operational tools for language modelling and the backbones of data-driven NLP methods. Yet, they seem inadequate when unseen MWEs are at stake, as these unseen ones may well be characterised by lexical combinations or syntactic structures that did not occur in annotated corpora and are thus hard to be identified automatically. Therefore, linking corpora and lexica would be beneficial for the robust MWE identification (Savary et al., 2019b). As of now, corpora and lexica remain to a great extent disconnected, with a few exceptions (Odijk, 2013; Markantonatou et al., 2019; Autelli, 2020) in which examples are extracted from corpora and added to the lexicon to illustrate the use of the MWEs.

Our goal is to design a lexicon-corpus interface that leverages MWE identification cross-linguistically. Three are the major challenges: (a) the harmonisation of corpora and lexica by also accounting for universality and diversity, (b) the efficient encoding of MWEs of all grammatical categories cross-linguistically, and (c) the adoption of the appropriate mechanisms and tools for linking lexica and corpora. Our work has been organised along three axes:

i capturing universality via cross-language unification of lexical features,

ii designing a lexicon-corpus interface usable for several languages, and

iii proof-of-concept encoding of MWEs based on the outcomes of (i) and (ii).

3. MWEs in computational lexica: state-of-the-art

In order to overview the state-of-the-art in the development of computational lexica of MWEs, we collected information about resources in a structured and systematic way, focusing on those published since 2016, as those published before this year were included in the survey performed within the COST Action PARSEME (Losnegaard et al., 2016). We have retrieved information for 75 resources from the following sources: European Language Grid repository, using the keyword "expressions" in the category Lexical/Conceptual resources; the ACL Anthology, in which we also used a keyword search (*multiword, idiom, phraseology*, etc.); the Phraseology and Multiword Expressions book series published by Language Science Press, and Europhras conferences, which were manually examined.

The data was harmonised aiming at a uniform and comprehensive description of the identified resources. It was organized in the following sections: General information (general or dedicated lexicon, mono- or multi-lingual), Corpus (in cases where the resource is related to a corpus), Resource (size, owner, licensing), Lemma & Representations (whether the resource provides information about the "lemma" of the MWE and its morphosyntactic properties), Syntax (details about syntactic information about the MWEs), Semantics (whether the resource provides semantic information about the MWEs and of what type) and References (major publication(s) about the identified resource).

The general picture obtained so far shows that:

- 72% of the resources are aimed for NLP use.
- More than 40 languages and dialects are represented, mostly Indoeuropean ones.
- 70.7% of the resources are monolingual, 18.7% bilingual and 10.6% multilingual.
- Most datasets were acquired manually or semi-automatically (automatically collected and manually verified).
- Only 24% of the resources are linked to a corpus and 12% are linked to other resources. The resources are usually linked to small purpose-built corpora. Usage examples are sometimes collected from a large representative corpus (without linking to the corpus).
- With regards to the encoded information, 45% of the resources provide comprehensive description of MWEs (including morphological, syntactic and semantic information). Semantic information, in particular, is extremely diverse.

The survey on MWE lexica raises several significant questions related to handling universality and diversity. First, most resources assume that a MWE entry is the coupling of a "lemma" form with a meaning. The definition of the "lemma" form is an open issue (see also section 4). In addition, often MWEs have "lemma" variants due not to grammatical phenomena but, for instance, to mutually exclusive choices of functional words or to the optionality of articles, and still, all these forms correspond to one meaning. It has been up to each resource's authors to decide which of these forms represents the MWE as its "lemma form" and how all these forms are related among them. As a result, different resources encode essentially the same MWE under different entries, as shown in Ex. 1 for Greek. Guidelines are needed even at this elementary level.

(1) [el]

vazo	(ti)	thilia	sto	lemo	kapiou
put	(the)	noose	to.ADP.the	neck	someone.GEN

vazo	(ti)	thilia	giro	apo	to	lemo	kapiou
put	(the)	noose	around.ADV	from.ADP	the	neck	someone.GEN

'to force someone to be involved in an unpleasant situation'

Second, various resources encode a different set of morphosyntactic and semantic features, in some cases with different degree of granularity, which poses a problem for their combined use and mutual enrichment. Guidelines handling the diversity among languages, in terms of morphological and syntactic properties of MWEs would facilitate their uniform representation and boost their NLP applications.

4. Universality: on cracking hard nuts

The notion of "word" is central to UD, but it is hard to define it in the context of the various typologically diverse languages. Thus, as a starting point of comparison, the strategy proposed by Haspelmath (2023) is followed. According to Haspelmath, 'A word is (i) a free morph, or (ii) a clitic, or (iii) a root or a compound possibly augmented by nonrequired affixes and augmented by required affixes if there are any.'. He also defines all the terms that constitute this definition: a free morph, a clitic, roots of various kinds, a compound, required/nonrequired affixes. Even with this typologically friendly approach, there exist a number challenges in a cross-lingual context. The main ones are: demarcation of clitics (words) vs. affixes (non-words), analysis of the compounds, marking the places of contraction splits.

For better modeling of data on the word level, a survey was conducted with Haspelmath's criteria. Responses for 43 languages were received. Based on that, a second version of the survey is being prepared that will allow for better comparison among language-specific properties. This new survey will target UD and non-UD languages and ask for examples of all of Haspelmath's word types that occur in the language. For UD languages, it will also ask for divergences between Haspelmath words and treebank words.

Although lemmatization may seem a very straightforward process and a solved task, this is quite misleading, because there exists a number of problems both in the lemmatization of words and in that of MWEs. The guidelines from UD and PARSEME say relatively little about lemmatization from a linguistic point of view. The focus there has been predominantly on tokenization and morphosyntactic analysis before the application of various linguistic tests and proposed classifications. For example, the relation between a token and a word is discussed in Savary et al. (2018): a token coincides with a word, several tokens constitute a multiword and one multiword contains several tokens. In UD the following is said: "The LEMMA field should contain the canonical or base form of the word, which is the form typically found in dictionaries. If a language is agglutinative, this is typically the form with no inflectional affixes; in fusional languages, the lemma is usually the result of a language-particular convention. If the lemma is not available, an underscore ("_") can be used to indicate its absence.". It means that the majority of decisions are left within the hands of treebank providers. Also, the guidelines say that "Except perhaps in rare cases of suppletion, one form should be chosen as the lemma of a verb, noun, determiner, or pronoun paradigm".

Various frameworks and annotation schemes apply different strategies to lemmatization and identify various issues. For example, Mambrini and Passarotti (2019) point to the following challenges in relation to Latin: the graphical representation, the spelling, the word ending, the representative paradigmatic slot, the homographic lemmas, the ambiguity in choosing the lemma, for example for participles that are hybrid forms and can be viewed either as verb forms by origin or as adjectives in some of their usages. The same holds for the deadjectival adverbs that can be viewed as part of the adjective paradigm or have their own lemmas. In (Mubarak, 2018) it is shown that the lemmatization task is quite complex for Arabic. The main linguistic problem is the mismatch between a word with a diacritic and its context (e.g. nouns and adjectives).

We outline only some of the challenges here. They refer to the issues of selecting the right form as a lemma, the existence of two options, the graphic representation varieties, the spelling specifics, the relation between inflection and derivation, the relation between orthographic words, their meaning and their spelling. The presented examples below feature some frequent lemma assigning problems across annotation schemes – within a single language and among languages. The list is not exhaustive, but it reflects the situation in many other languages and frameworks. Since this task is work in progress, the plan is to study the lemmatization decisions in the various UD treebanks and in PARSEME corpora as being already very multilingual and as sources of integration of these two frameworks and data, and also beyond them – through investigating papers on different language families, as well as through questionnaires.

Lemmatization challenges of some words and tokens

- *Pronouns*. In some languages (like Bulgarian, Czech, Maltese) there are short and long forms of some pronouns (e.g. personal), or strong and weak ones (like in Greek and Italian). Thus, the following possibilities for lemmatization exist for the short 3rd person pronouns in Czech, for example: a) the lemma equals the wordform itself (`[cs]`: *ho*-3P.MASC.SG.ACC.SHORT 'him'), b) the lemma goes to the long 3rd person form ([cs]: něho-3P.MASC.SG.ACC.SHORT 'him'), c) the lemma goes to the nominative, masculine, 3rd person form (`[cs]`: *on*-3P.MASC.SG.NOM 'he'), while in d) the lemma is the pronoun in 1st person, singular, nominative as the less marked form (`[cs]`: *já*-1P.SG.NOM 'I'). Thus, different strategies can be applied with varying depth until reaching the lemma.

- *Doublets*. There are doublet verbs that share the same paradigm. For example, the same lemma verb with two different endings (`[bg]`: *zna-m* and *zna-ya* (lit. *know-I*) 'to know'); or the same lemma adjective with two different variants (`[bg]`: *sasht* and *sashti* 'same-M.SG'). Thus, one of the doublets might be selected as representative, but it is sometimes hard to make such a selection.

- *Numbers*. In text data, numbers can occur as words or as digits. Should both representations of the same number have the same lemma? And if so, then which one?

- *Negated words*. This problem relates also to graphic conventions. In some languages, the negation of a word is written together, for example – as a prefix. In Bulgarian, this holds for the nominals, in Czech this holds also for verbs, while in Romanian it holds for some nominals and for three out of the four non-finite forms of a verb (only for participle, supine and gerund, but not for infinitive). Should the lemma of the negated word be its positive counterpart (meaning that negation is treated rather like inflection than derivation)?

- *Diminutives*. Although the process of making diminutives is derivational, it is still not clear whether the lemma of the word should be the diminutive or the original word. According to the current UD guidelines, the lemma does not remove derivational morphology. If such a strategy is followed, the lemma should be the diminutive. However, if most of the diminutives are not part of the dictionary, then there might be problems during the next NLP processing tasks.

Lemmatization challenges of some MWEs

- *Compounding*. In many languages, a compound (traditionally a word with (at least) two roots) can be written differently: as two words, as one word or with a hyphen. Compare in Bulgarian the double spelling: *biznes plan* (two words) and *biznesplan* (one word), in English *business plan* (two words) and in German *Businessplan* (one word). A problem arises when trying to offer a uniform analysis of these compounds within a language and across languages.

- *(Quasi)reflexive verbs*. Even within one language family like the Slavic languages, the quasi-reflexive particle can be either a separate word (`[bg]`: *smeya se*, `[cs]`: *smát se* 'to laugh') or part of the word (`[uk]`: *smijatysja* 'to laugh'). The reflexive pronouns are part of the word also in some Romance languages (`[es]`: *lavarse* 'to wash oneself') and not in others (`[ro]`: *se spăla* 'to wash oneself'), but in the non-reflexive meaning they lose this clitic (*lavar* 'to wash something/someone'). The question is whether the lemma is defined within each language/language family on formal criteria, or there might be possibilities to create some cross-linguistic strategies.

5. Linking MWE lexicon entries with their occurrences in corpora

Publishing language resources as Linked Data enhances accessibility, interoperability, semantic enrichment, community collaboration, and the promotion of open science. These contribute to the advancement of linguistic research, language technology, and cross-disciplinary insights.

Analyzing unique language patterns across different languages can benefit from sharing aligned and annotated corpus data in a format that complies with community standards like the NLP Interchange

Format (NIF) (Hellmann et al., 2012, 2013) and CoNLL-RDF (Chiarcos and Fäth, 2017; Chiarcos and Glaser, 2020). CoNLL-RDF is a simplified version of NIF that aligns with tab-separated formats, such as CoNLL, CoNLL-U for Universal Dependencies, and Parseme-TSV for PARSEME.

Working towards the objective of designing a lexicon-corpus interface and prove its functionality, we will expand the existing ELEXIS-WSD Parallel Sense-Annotated Corpus (Martelli et al., 2023). Currently at version 1.1, it can be accessed from the CLARIN.SI repository[4] and contains 2,024 sentences across 10 languages, along with a sense repository for each language. The expansion of the corpus will involve adding new languages (Krstev et al., 2024) and upgrading the annotation to enable linking MWE lexicon entries with their occurrences in the corpora.

Moreover, these resources should also be published as Linked Data (using NIF) to facilitate linking with the sense repository of the corpus. For the ELEXIS dictionary data, the OntoLex vocabulary[5], a widely used community standard for machine-readable lexical resources in the context of RDF, Linked Data, and Semantic Web technologies (McCrae et al., 2017), will be considered, as it is currently the foundation for the majority of lexical data available on the web of data.

Apart from the core module **Lemon** with general data structures, OntoLex modules relevant to MWEs include the module for the internal structure and combinatory semantics of MWEs **Decomp**, and MWE morphology **Morph** module. The new module for Frequency, Attestations, and Corpus-based Information (**FrAC**)[6] (Chiarcos et al., 2022a,b) supports linking lexica with corpora in many aspects of information relevant to the joint work with corpora and dictionaries. **Lexicog** (Bosque-Gil et al., 2019) is a module for lexicography that addresses structures and annotations commonly found in lexicography. It is designed to operate in combination with OntoLex for the representation of dictionaries and any other linguistic resource containing lexicographic data.

An attempt at leveraging Linked Data, NIF, and CoNLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora is reported in the literature and could be followed (Stanković et al., 2024).

6. Proof-of-concept lexical encoding of MWEs

Taking the above into consideration, a proof-of-concept lexical encoding of MWEs in NLP lexica, that also maintains the lexicon-corpus interface, should minimally abide by the following requirements:

- a definition of the notion of "word" that is as universal as possible,
- a shared understanding of MWEs that can be annotated in corpora and then linked with lexicon entries (both the MWE as a whole and its components), including all types of MWEs (not only nominal and verbal),
- centralised guidelines for lexicon encoding regarding, i.e., the notions of lemma, canonical form, lexical features, etc.,
- a uniform representation of the syntactic properties of MWEs, and
- tools and mechanisms for linking MWE entries with their occurrences in corpora.

7. Conclusion

In an effort to create an ecosystem of interlinked MWE-dedicated lexica and annotated corpora, with an eye to universality and accommodating the languages specificities, we have already painted the current landscape of this field and are striving to find solutions for cracking the hard nuts (syntactic word definition, word and MWE lemmatization, lexical features, etc.) and to create guidelines for MWE lexicographic description. Development of linguistic resources for various languages in a harmonized way and their interlinking using standardization methods can only lead to the progress of language technology, as well as serve as a model for low-resourced languages in their endeavour to catch up with domain's evolution, speeding this process due to the benefits that Linked Data can offer (Bosque-Gil et al., 2022).

8. Acknowledgments

This paper is funded by the CA21167 COST Action UniDive, supported by COST (European Cooperation in Science and Technology).

9. Bibliographical References

Erica Autelli. 2020. *Phrasemes in Genoese and Genoese-Italian lexicography*, page 101–127.

[4] https://www.clarin.si/repository/xmlui/handle/11356/1842

[5] https://www.w3.org/2016/05/ontolex

[6] The current draft version of the FrAC specification is found under https://github.com/ontolex/frequency-attestation-corpus-information/

University of Białystok Publishing House., Białystok.

Julia Bosque-Gil, Dorielle Lonke, I Kernerman, and J Gracia. 2019. Validating the ontolex-lemon lexicography module with k dictionaries"multilingual data. In *Electron. lexicogr. 21st cent., Proc. eLex conf.*, ART-2019-123124.

Julia Bosque-Gil, Verginica Barbu Mititelu, Hugo Gonçalo Oliveira, Maxim Ionov, Jorge Gracia, Liudmila Rychkova, Giedre Valunaite Oleskeviciene, Christian Chiarcos, Thierry Declerck, and Milan Dojchinovski. 2022. Balancing the digital presence of languages in and for technological development, a policy brief on the inclusion of data of under-resourced languages into the linked data cloud.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027.

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1*, pages 74–88. Springer.

Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Besim Kabashi, Fahad Khan, and Ciprian-Octavian Truică. 2022b. Modelling Collocations in OntoLex-FrAC. In *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*, pages 10–18.

Christian Chiarcos and Luis Glaser. 2020. A tree extension for CoNLL-RDF. In *Proceedings of the 12th LREC*, pages 7161–7169.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Martelli Federico, Navigli Roberto, Krek Simon, Kallas Jelena, Gantar Polona, Veronika Lipp, Tamás Váradi, András Győrffy, and László Simon. 2021. Designing the elexis parallel sense-annotated dataset in 10 european languages. In *Proceedings of the eLex 2021 conference*, pages 377–395. Lexical Computing.

Martin Haspelmath. 2023. Defining the word. *WORD*, 69(3):283–297.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating nlp using linked data. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12*, pages 98–113. Springer.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Marcus Nitzschke. 2012. NIF Combinator: Combining NLP Tool Output. In *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, 2012. Proceedings 18*, pages 446–449. Springer.

Cvetana Krstev, Ranka Stanković, and Aleksandra Marković. 2024. Towards the semantic annotation of sr-elexis corpus: Insights into multiword expressions and named entities. In *Proc. of Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)*.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. PARSEME survey on MWE resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2299–2306, Portorož, Slovenia. European Language Resources Association (ELRA).

Francesco Mambrini and Marco Passarotti. 2019. Harmonizing different lemmatization strategies for building a knowledge base of linguistic resources for Latin. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 71–80, Florence, Italy. Association for Computational Linguistics.

Stella Markantonatou, Panagiotis Minos, George Zakis, Vassiliki Moutzouri, and Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. In *Proceedings of Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Workshop at ACL 2019*, pages 130–134, Florence, Italy. Association for Computational Linguistics (ACL).

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: Development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Hamdy Mubarak. 2018. Build fast and accurate lemmatization for Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jan Odijk. 2013. *Identification and lexical representation of multiword expressions*, pages 201–217. Springer Berlin Heidelberg, Berlin, Heidelberg.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten Van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke Der, Behrang Qasemi Zadeh, Carlos Ramisch, and Veronika Vincze. 2018. *PARSEME multilingual corpus of verbal multiword expressions*.

Agata Savary, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019a. Literal occurrences of multiword expressions: Rare birds that cause a stir. 112:5–54.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019b. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.

Ranka Stanković, Christian Chiarcos, and Milica Ikonić Nešić. 2024. Leveraging Linked Data, NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora. In *Book of Abstracts of the UniDive 2nd general meeting, 8-10 February 2024, Naples*.

10. Language Resource References

Martelli, Federico and Navigli, Roberto and Krek, Simon and Kallas, Jelena and Gantar, Polona and Koeva, Svetla and Nimb, Sanni and Sandford Pedersen, Bolette and Olsen, Sussi and Langemets, Margit and Koppel, Kristina and Üksik, Tiiu and Dobrovoljc, Kaja and Ureña-Ruiz, Rafael and Sancho-Sánchez, José-Luis and Lipp, Veronika and Váradi, Tamás and Győrffy, András and Simon, László and Quochi, Valeria and Monachini, Monica and Frontini, Francesca and Tiberius, Carole and Tempelaars, Rob and Costa, Rute and Salgado, Ana and Čibej, Jaka and Munda, Tina and Kosem, Iztok and Roblek, Rebeka and Kamenšek, Urška and Zaranšek, Petra and Zgaga, Karolina and Ponikvar, Primož and Terčon, Luka and Jensen, Jonas and Flörke, Ida and Lorentzen, Henrik and Troelsgård, Thomas and Blagoeva, Diana and Hristov, Dimitar and Kolkovska, Sia. 2023. *Parallel sense-annotated corpus ELEXIS-WSD 1.1*. Jožef Stefan Institute. Slovenian language resource repository CLARIN.SI.

Annotation of Multiword Expressions in the SUK 1.0 Training Corpus of Slovene: Lessons Learned and Future Steps

Jaka Čibej, Polona Gantar, Mija Bon
Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, 1000 Ljubljana, Slovenia
{jaka.cibej, apolonija.gantar, mija.bon}@ff.uni-lj.si

Abstract

Recent progress within the UniDive COST Action on the compilation of universal guidelines for the annotation of non-verbal multiword expressions (MWEs) has provided an opportunity to improve and expand the work previously done within the PARSEME COST Action on the annotation of verbal multiword expressions in the SUK 1.0 Training Corpus of Slovene. A segment of the training corpus had already been annotated with verbal MWEs during PARSEME. As a follow-up and part of the New Grammar of Modern Standard Slovene (NSSSS) project, the same segment was annotated with non-verbal MWEs, resulting in approximately $6,500$ sentences annotated by at least three annotators (described in Gantar et al., 2019). Since then, the entire SUK 1.0 was also manually annotated with UD-part-of-speech tags. In the paper, we present an analysis of the MWE annotations exported from the corpus along with their part-of-speech structures through the lens of Universal Dependencies. We discuss the usefulness of the data in terms of potential insight for the further compilation and fine-tuning of guidelines particularly for non-verbal MWEs, and conclude with our plans for future work.

Keywords: multiword expressions, Universal Dependencies, Slovene

1. Introduction

Slovene was one of the languages involved in the PARSEME COST Action [1]. As part of the activities, $11,411$ sentences (approx. 41 %) of the ssj500k 2.1 Slovene Training Corpus (Krek et al., 2018)[2] were annotated with verbal MWEs (Gantar et al., 2017) categorized according to the PARSEME guidelines and MWE-tests (Savary et al., 2018). Work on Slovene MWEs within the same corpus then continued after the conclusion of PARSEME within the national project titled *New Grammar of Contemporary Standard Slovene: Sources and Methods*[3], during which non-verbal MWE annotations were added to $6,500$ sentences (a subset of the $11,411$ sentences annotated within PARSEME). Non-verbal MWEs were annotated (the process is described in more detail in (Gantar et al., 2019)) according to a set of guidelines designed primarily from the point of view of inclusion of MWEs in dictionaries, while the categorization principles followed the definitions used in the compilation of Slovene Lexical Database (Gantar and Krek, 2011) and the Digital Dictionary Database of Slovene (Kosem et al., 2021). However, the annotations have so far not been included in the SUK 1.0 corpus, pending an additional curation and resolution of crucial questions, mainly which of the annotated spans should be considered MWEs, particularly with regard to multiword combinations with varying levels of terminologicalness.

Recent advances within the UniDive COST Action[4], which among its tasks (specifically in Task 1.2) also includes the extension of the PARSEME verbal MWE annotation guidelines[5] with non-verbal MWEs, have provided an opportunity to continue the work already done on Slovene MWE annotations in the SUK 1.0 corpus within other projects, as well as to compare our own MWE-categorization with the one adopted within UniDive. At the time of writing this paper, the UniDive non-verbal MWE annotation guidelines contain no examples of Slovene MWEs, and a discussion is still underway. In addition to these examples, the lessons from the annotation of the SUK 1.0 corpus may provide a number of valuable insights during the initial phase of uni-

[1] *Parsing and multi-word expressions. Towards linguistic precision and computational efficiency in natural language processing*, IC1207 COST Action, 2013-2017: https://typo.uni-konstanz.de/parseme/

[2] Since then, the ssj500k training corpus was extended with several other datasets and underwent a rebranding, now being called the SUK 1.0 Training Corpus of Slovene (Arhar Holdt et al., 2022). In this paper, we refer to it using the new name unless we specifically refer to an older version. The SUK 1.0 corpus consists mostly of newspaper texts, magazines, and internet texts, with a small percentage of fiction and non-fiction.

[3] New Grammar of Contemporary Standard Slovene - project website: https://slovnica.ijs.si/?lang=en

[4] *Universality, Diversity and Idiosyncrasy in Language Technology*, CA21167 COST Action, 2022-2026: https://unidive.lisn.upsaclay.fr/

[5] PARSEME Annotation guidelines 1.3 - https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 154–162. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

fying the PARSEME annotation scheme with Universal Dependencies (Savary et al., 2023). While the data only covers Slovene, its advantage is that several statistical calculations were made based on the annotations, including for example the scope of MWE annotation and length overlap, as well as inter-annotator agreement (each sentence was annotated by at least three annotators). In the paper, we discuss the annotated MWEs and focus predominantly on the points of disagreement and lessons learned that may prove useful for the compilation of MWE annotation guidelines within UniDive. The paper is structured as follows: we first provide a short overview of related work on MWEs (Section 2) and describe the data on annotated MWEs exported from the SUK 1.0 corpus (Section 3), then provide an analysis (Section 4). We conclude the paper with a discussion on the usefulness of the data within UniDive and a list of potential future steps in our work.

2. Related Work

MWEs still pose a problem for NLP tools such as machine translation systems, word sense disambiguation, or computational lexicography (e.g. MWE detection in corpora). A number of endeavors have been undertaken to provide training or evaluation datasets annotated with MWEs, both monolingual (Adalı et al., 2016 for Turkish; Candito et al., 2020 for French; Kato et al., 2018 and Schneider et al., 2014 for English; Mohamed et al., 2022 for Arabic; Souza and Freitas, 2023 for Portuguese) and multilingual (Monti et al., 2015; Han et al., 2020; Savary et al., 2018).

So far, no Slovene manually annotated corpus includes comprehensive and systematic annotations of MWEs; aside from the already mentioned PARSEME verbal MWE annotations in the ssj500k 2.1 Training Corpus (Gantar et al., 2017) which also serves as the Slovene UD Treebank, a small dataset for the automatic detection of idiomatic expressions has also been made by Škvorc et al. (2022) in order to facilitate idiomatic expression extraction using contextual embeddings. There is also the Slovene subcorpus of the ELEXIS-WSD Parallel Sense-Annotated Corpus (Martelli et al., 2021); however, MWEs within the corpus have not been categorized and only their spans have been annotated, while the corpus itself was primarily compiled for word sense disambiguation focused on single word units.

The first step toward extending the SUK 1.0 corpus with comprehensive MWE annotations was made by (Gantar et al., 2019) by conducting an experimental annotation campaign to identify potential MWE candidates. We discuss the results in the following sections.

3. Data Description

The annotation process and the typology used to annotate MWEs in SUK 1.0 was described in detail by (Gantar et al., 2019), so we only provide a brief overview here. The main goal of the task was to annotate non-verbal multiword expressions according to a typology that defines two main subgroups of MWEs[6]: (a) *lexical units*, which require an explanation (due to them being characterized by a certain degree of semantic non-compositionality), and (b) *lexico-grammatical units*, which are semantically relatively transparent (they complement or disambiguate the sense description of a headword (e.g. collocations) or they play a role of syntactic connectors or discourse organizers in language[7]).

Multiword lexical units are further divided into *fixed expressions* (which typically cover terminological expressions such as *črna luknja* 'black hole' in the sense of an astronomic phenomenon) and *phraseological units* (which typically express a metaphorical or pragmatic meaning, such as *princ na belem konju* lit. 'prince on a white horse'; 'knight in shining armor').

Lexico-grammatical units, on the other hand, consist of *collocations* (not included in the annotation task as they are regarded as semantically transparent (Atkins and Rundell, 2008) and can be automatically extracted from corpora using several criteria such as structure and statistical co-occurrence) and *syntactic combinations* (which have no lexical meaning, but are nevertheless relevant for lexicographic description because they act as adverbials, sentence connectors, and discourse organizers; such as *v skladu z* 'in accordance with').

The annotators were thus tasked with annotating MWEs as either phraseological units (PU), fixed expressions (FE), or syntactic combinations (SC). It should be noted that this is a parallel categorization, so the existing verbal MWEs annotated within PARSEME were also assigned additional categories according to this system. In this paper, we focus on the annotated UD POS-structures and patterns, not the categorization according to our own typology; more detailed results of the categorization were already presented in Gantar et al.,

[6] This categorization follows the organization of language data in the Digital Dictionary Database of Slovene (Kosem et al., 2021), where the main criterion to distinguish different types of MWEs depends on whether the MWE is a semantically independent or dependent unit.

[7] In retrospect, it should be mentioned that the decision to explicitly categorize discourse organizers as lexico-grammatical units caused some disagreement during annotation; if a discourse organizer (such as *v bistvu* 'in fact', 'actually') requires a semantic explanation and plays a pragmatic role in the sentence that needs to be explained in a dictionary, it should be categorized as a phraseological unit (PU), which falls under lexical units.

2019.

The annotation resulted in a total of 15,727 MWE annotations in the first 6,500 sentences of the SUK 1.0 corpus. Each sentence was annotated by at least 3 annotators (see Gantar et al., 2019), so a potential MWE-candidate within an individual sentence has up to three annotations (depending on whether the annotator identified the span as a MWE). For instance, in the sentence below, two annotators identified one MWE candidate and each provided an annotation; one annotated *v nasprotju* (lit. 'in contradiction') while the other annotated *v nasprotju s* (lit. 'in contradiction with').

sl Toda [[v nasprotju] s] svojimi sorodniki sodijo kaneloni (cannello = cevka) šele slabih sto let k italijanski testeninski klasiki.

en But contrary to their relatives, cannelloni (cannello = tube) have been a part of the Italian pasta classics for less than one hundred years.

A total of 8,864 MWE candidates were annotated in the corpus, consisting of 6,385 different potential MWEs.[8]

Since the annotations were made, a section of the SUK 1.0 corpus was also manually annotated with UD-part-of-speech tags, UD dependency relations, and named entities (see Arhar Holdt et al., 2023); this includes the 6,500 sentences annotated with both verbal and non-verbal MWEs, which enables us to export MWE annotations along with UD part-of-speech tags, dependency relations, and named entities, and observe potential patterns as well as points of potential disagreement. We provide a thorough analysis in Section 4 below.

4. Analysis

As shown in Table 1, the MWE candidates were annotated by 10 annotators; two of which (A and B) were reference annotators involved in the compilation of the annotation guidelines. The rest were students of linguistics at the University of Ljubljana. The distribution of annotations and the average number of MWE annotations per sentence shows that most of the annotators annotated MWEs similarly frequently to the reference annotators (approx. 0.5–0.6 MWEs per sentence), with two outliers, who were either too liberal (annotator I) or too strict (annotator J).

Out of 8,864 annotated MWE candidates, 5,023 (56.67%) were assigned a single annotation, 2,103 (23.73%) two annotations, and 1,738 (19.61%) three or more annotations. As shown in Table 2, a large portion of single annotations (almost 40%) were made by the most liberal annotator (I), but a significant percentage was provided by other annotators as well, including one of the reference annotators (B, with approx. 11%).[9] As the identification of MWEs is a difficult task, a certain degree of disagreement is to be expected. In the following subsections, we further analyze the annotations in order to discover any recurring misinterpretations that could point to potential gaps in the annotation guidelines.

Ann.	MWEs	Sent.	%	MWE/Sent.
A	292	500	1.86%	0.584
B	3,111	6,500	19.86%	0.479
C	1,742	2,000	11.12%	0.871
D	1,716	2,000	10.95%	0.858
E	1,124	2,000	7.17%	0.562
F	1,367	2,000	8.73%	0.683
G	903	2,000	5.76%	0.452
H	1,467	2,000	9.36%	0.734
I	3,563	2,000	22.74%	1.782
J	382	2,000	2.44%	0.191

Table 1: Table of MWE-annotations showing individual annotators, the number of MWE-annotations they made in the corpus, the number of all sentences annotated by them, the percentage of all annotations made, and the number of MWEs per sentence.

Ann.	Single cand.	%
I	1,953	38.86%
D	696	13.86%
C	601	11.96%
B	547	10.89%
H	380	7.57%
F	313	6.23%
G	307	6.11%
E	126	2.51%
J	91	1.81%
A	9	0.18%

Table 2: Distribution of single-annotation MWE candidates across annotators.

4.1. Part-of-Speech Structure

Based on the annotated tokens and their UD part-of-speech tags, the annotated MWE candidates cover 920 different structures, with the top 17 accounting for approx. 65% of all annotations (see Table 3). Each of these covers more than 1% of the annotations, while the other categories cover less. The majority of the annotations are non-verbal, with verbs

[8] The 6,385 different candidates were counted based on the alphabetical combinations of lemmas within annotated spans.

[9] A more detailed calculation of MWEs missed or intentionally left unannotated by individual annotators can be made once the final annotations have been encoded in the corpus.

Structure	MWE Ann.	%
ADJ NOUN	4,550	29.04%
ADP NOUN	2,053	13.10%
ADP DET	401	2.56%
NOUN NOUN	391	2.50%
VERB ADP NOUN	360	2.30%
PART AUX	353	2.25%
ADP DET NOUN	298	1.90%
PART ADV	228	1.46%
ADJ ADJ NOUN	224	1.43%
ADP ADJ NOUN	214	1.37%
NOUN ADP NOUN	187	1.19%
VERB NOUN	186	1.19%
ADP ADJ	174	1.11%
DET SCONJ	174	1.11%
ADV SCONJ	171	1.09%
ADP NOUN ADP	168	1.07%
ADP ADP	165	1.05%
Other	5,658	35.98%

Table 3: Distribution of MWE annotations based on their UD part-of-speech structure.

featured in only two of the most frequent categories (VERB ADP NOUN and VERB NOUN). The most frequent part-of-speech structure is ADJ NOUN (e.g. *sodni postopek*, 'judicial process', *vozniško dovoljenje*, 'driver's license'), covering almost a third of all annotations, and ADP NOUN (e.g. *v celoti*, lit. 'in whole', 'entirely'; *pred časom*, lit. 'before time', 'some time ago').

We analyzed the distribution of the part-of-speech structures in terms of how prone they were to single annotations in order to check whether any structure is more problematic for MWE identification. Table 4 shows the 10 most frequent part-of-speech structures that are also more typical of single annotations compared to all annotations (i.e. according to the ratio in the last column, they are more likely to be annotated by just a single annotator and less likely to be annotated multiple times).

An analysis of the single annotation examples with these structures reveals a number of problematic groups, particularly within structures with a nominal distribution (e.g. NOUN NOUN, NOUN ADP NOUN). First, there are terminological candidates that may be somewhat compositional, but have a specific meaning within a certain field (e.g. *omejevalnik vrtljajev* 'rev limiter', *raziskave tržišča* 'market research', *vitamin C*, 'vitamin C'). In some cases, the annotated spans are collocations that are semantically transparent, but very typical (e.g. *kraj zločina*, lit. 'place of the crime', 'scene of the crime'; *balzam za ustnice*, 'lip balm'). Secondly, some spans denote titles or functions (e.g. *poveljnik straže*, 'captain of the guard'; *hranilec družine*, lit. 'feeder of the family', 'family provider') or even members of an association or organization (e.g. *sestre usmiljenke*, 'Sisters of Mercy'), which should be treated more as named entities despite not being capitalized. Similarly, the third problematic group contains spans that can be interpreted as named entities, but that is not entirely clear when the span is spelled with no capitalization and the context is somewhat ambiguous whether the examples refer to concrete instances or a general concept (e.g. *liga prvakov*, 'league of champions'; *ministrstvo za finance*, 'ministry of finance'). In addition, examples contain phrases in which one of the components exhibits a metaphoric meaning - e.g. *gostja večera*, 'guest of the evening' in the sense of 'the guest of tonight's show'), which prompts the annotator to treat the span as non-compositional.

Next, there are several grammatical constructions that were mistakenly annotated as multiword expressions, such as combinations of prepositions and relative pronouns (ADP DET; *v kateri* 'in which', *po kateri* 'after which', *h kateri* 'to which'); some of the annotators probably annotated these because *kateri* as a relative pronoun only occurs next to prepositions, so they treated both components as a single unit. Similarly, sequences of prepositions and demonstrative pronouns (*glede tega* 'regarding this', *iz tega* 'from this') occurring in a very vague context could have prompted to treat them as non-compositional, as in the example below:

sl Država **s tem** priznava, da je prostovoljnih vojakov premalo, če ne kar nič.

en **With this**, the State recognizes that there are too few voluntary soldiers, if any.

Interestingly, some candidates with similar part-of-speech structures (either ADP DET or ADP PRON) do represent legitimate MWEs (e.g. *po svoje*, 'in its own way'; *pri nas*, lit. 'at us', 'in our country'), but were only annotated once, which indicates that expressions containing mostly closed-class parts-of-speech (which frequently constitute syntactic combinations according to our typology) should be described in more detail in the guidelines, with additional negative examples. Before manually annotating additional sentences in the corpus, a more targeted approach could be taken by extracting n-grams with problematic closed-class structures and creating a list of all syntactic combinations discovered this way (e.g. two-part connectors such as *ne samo A, temveč tudi B* 'not only A, but also B').

Table 5, on the other hand, shows the part-of-speech structures that were more likely annotated by multiple annotators (3 or more). The most frequent structure, VERB ADP NOUN (e.g. *vzeti pod drobnogled*, lit. 'take [sth] under the microscope', 'to take under scrutiny'), was frequently and consistently annotated because it contains verbal MWEs previously annotated with PARSEME categories

Struct.	Sin.	% (Sin.)	% (All)	Ratio
NOUN NOUN	195	3.88%	2.5%	1.55
ADP DET	172	3.42%	2.56%	1.34
PART ADV	88	1.75%	1.46%	1.2
PROPN	72	1.43%	0.63%	2.27
NOUN ADP NOUN	71	1.41%	1.19%	1.18
ADP PRON	68	1.35%	0.69%	1.96
VERB ADV	50	1.0%	0.54%	1.85
ADV CCONJ	46	0.92%	0.43%	2.14
SCONJ AUX	42	0.84%	0.53%	1.58
CCONJ PART	37	0.74%	0.29%	2.55

Table 4: Comparison of the distribution of part-of-speech structures between single annotations and all annotations (10 most frequent structures that are also most typical of single annotations). The columns show the number of single annotations within the structure, the percentage that structure covers within single annotations, the percentage it covers in all annotations, and the ratio between percentages.

(which the annotators followed). The second structure (PART AUX) contains just one MWE, *naj bi*, which is a very crystallized expression used in the sense of 'is said to', and was mentioned in the guidelines as a good example of a syntactic combination. Among the more intuitive structures are ADP DET NOUN (*po vsej verjetnosti*, 'in all likelihood'; *do te mere*, 'to such a degree'), ADP NOUN ADP (*v zvezi z*, lit. 'in connection with', 'with regard to', *v skladu z*, 'in accordance with'), and ADP ADJ (*med drugim*, 'among other things'; *pred kratkim*, 'a short while ago'). Some structures confirm that generating a list of MWEs containing closed-class elements would be useful: for instance, ADP ADP (*od - do*, 'from - to'), NUM ADP (*eden od*, 'one of') and DET SCONJ (*več kot*, 'more than') were quite consistently annotated because they were listed in the guidelines. The same goes for abbreviations (X and X X, such as *itn.*, *in tako naprej*, 'and so on'; *t. i.*, *tako imenovani*, 'so-called'), which could also be extracted and included in a reference list.

The two most frequently annotated structures in general (ADJ NOUN and ADP NOUN) appear almost equally frequently in both the single annotations as well as multiple annotations. This is to be expected, as the difference between a MWE and, for instance, a collocation or a terminological candidate, is a question of semantic interpretation, particularly in the context of the guidelines used for this annotation task, which relied heavily on the annotator's interpretation on whether an annotated span would require a semantic or encyclopedic explanation in a (general) dictionary language resource.

Struct.	Mul.	% (Mul.)	% (All)	Ratio
VERB ADP NOUN	232	3.6%	2.3%	1.57
PART AUX	216	3.35%	2.25%	1.49
ADP DET NOUN	169	2.62%	1.9%	1.38
ADP NOUN ADP	127	1.97%	1.07%	1.84
NUM ADP	115	1.79%	0.99%	1.81
ADP ADP	113	1.75%	1.05%	1.67
DET SCONJ	112	1.74%	1.11%	1.57
ADP ADJ	108	1.68%	1.11%	1.51
X X	72	1.12%	0.68%	1.65
X	66	1.03%	0.54%	1.91

Table 5: Comparison of the distribution of part-of-speech structures between multiple annotations and all annotations (10 most frequent structures that are also most typical of multiple annotations).

4.2. Annotation Scope and Overlap

In this section, we analyze the degree to which the annotators agreed on the scope of the annotation of individual MWE candidates. Out of the $8,864$ annotated candidates, $5,023$ (56.67%) were annotated by a single annotator, while $3,841$ (43.33%) were assigned multiple annotations. Out of these $3,841$ candidates, $2,961$ (77.10%) exhibited complete overlap, meaning that all the annotators annotated the exact same elements in each case. The vast discrepancy between single annotations and the percentage of candidates with complete overlap indicates that while there is disagreement on whether a span is a MWE, in the majority of examples where a span is identified as a MWE by multiple annotators, they tend to agree on the elements included. Only 880 examples showed disagreement in annotation scope. For each candidate with incomplete overlap, we first aggregated all the annotated elements and identified the ones that differed between the annotations. In the example below, the MWE candidate was independently annotated four times (*Prav tako*, *tako kakor*, *Prav tako*, *Prav tako kakor*). Only the element *tako* (ADV) appears in all annotations, while *prav* (PART) and *kakor* (SCONJ) do not, so they are treated as differing elements.

sl **Prav tako** jasen **kakor** prejšnji, bilo je le nekoliko hladneje.

en **Just as** clear **as** the day before; it was only somewhat colder.

Table 6 shows the distribution of differing elements by part-of-speech. While adjectives and nouns are at the top of the list, prepositions (ADP), determiners (DET), pronouns (PRON), particles

UPOS	Nr.	%
ADJ	227	16.85%
NOUN	210	15.59%
ADP	172	12.77%
VERB	163	12.10%
DET	116	8.61%
AUX	116	8.61%
PRON	73	5.42%
PART	72	5.35%
ADV	62	4.60%
CCONJ	57	4.23%
SCONJ	56	4.16%
NUM	18	1.34%
PROPN	5	0.37%

Table 6: Frequencies and percentages of parts of speech causing disagreement in MWE scope annotation.

(PART) and conjunctions (SCONJ, CCONJ) account for more than 40% of all differing elements.

To identify potential recurring points of disagreement within specific part-of-speech structures, we also exported co-occurrences of differing structures from annotations with incomplete overlap. So for the example above (*prav tako kakor*), all the different structures were the following: *Prav tako*, PART ADV; *tako kakor*, ADV SCONJ, *Prav tako*, PART ADV; *Prav tako kakor*, PART ADV SCONJ. We counted all the possible combinations of two (excluding the ones with equal pairs) to obtain counts of the most frequently co-occurring structures. $4,063$ co-occurrences of differing structure pairs were counted and further analyzed; a selection of the most interesting pairs is shown in Table 7.

The examples in which an adjective was the contested element reveal some interesting insights: the ADJ ADJ NOUN - ADJ NOUN dilemma raises the issue of annotating potential nested MWEs (*varuh človekovih pravic*, 'human rights ombudsman' vs. *človekove pravice*, 'human rights'), as well as the issue of optional vs. obligatory elements in MWEs (e.g. *človeške pravice*, 'human rights', vs. *temeljne človeške pravice*, 'fundamental human rights'). This is similar to ADP ADJ NOUN - ADP NOUN (*po ocenah*, 'according to estimates' vs. *po prvih ocenah*, 'according to the first estimates'). While the guidelines provided instructions on how to treat some of the optional elements, they were mainly focused on the inclusion of verbs in examples such as *pisati na roko*, 'to write by hand'). As a general rule, however, each example was to be annotated individually based on how typical the syntactic environment of the identified MWE was, along with the relevant lexical elements. For further annotation, the treatment of these elements should be further specified in order to avoid disagreement.

Diff.	Str. Pair	Freq.
ADJ	ADJ ADJ NOUN - ADJ NOUN	208
ADJ	ADP ADJ NOUN - ADP NOUN	65
NOUN	ADJ NOUN - NOUN ADJ NOUN	79
NOUN	ADJ NOUN - NOUN ADP ADJ NOUN	23
VERB	ADP NOUN - VERB ADP NOUN	90
ADP	ADJ NOUN - ADP NOUN	62
ADP	ADJ NOUN - ADP ADJ NOUN	41
AUX	AUX VERB ADP NOUN - VERB ADP NOUN	24
AUX	AUX VERB NOUN - VERB NOUN	20
DET	ADP DET NOUN - ADP NOUN	91
PART	ADP NOUN - PART ADP NOUN	19
CCONJ	ADP DET - ADP DET CCONJ	35
NUM	ADP NOUN - ADP NUM NOUN	15

Table 7: Most frequent co-occurring structures within annotations with incomplete overlap. The first column denotes the differing element, the second the structure pair, and the third the frequency of co-occurrence.

When nouns are the differing element, the examples again show some discrepancy when it comes to potential nested MWEs (e.g. ADJ NOUN - NOUN ADJ NOUN; *ponudniki mobilnih signalov*, 'mobile signal providers' vs. mobilni signal, 'mobile signal'; *šef obveščevalne službe*, 'secret service director' vs. *obveščevalna služba*, 'secret service'; or ADJ NOUN - NOUN ADP ADJ NOUN; *rak na materničnem vratu*, lit. 'cancer on the uteral neck', 'cervical cancer' vs. *maternični vrat*, 'cervix'). The current annotation task did not include the annotation of nested MWEs, but the results show that the guidelines should be extended to address this topic and provide clearer instructions (either by allowing for nested annotations or by listing principles on how to determine the optimal scope of the MWE).

The examples with verbs as the differing element seem to indicate that the pool of available lexical candidates that can be substituted within a MWE affects the annotator's scope. For instance, the structure pair ADP NOUN - VERB ADP NOUN contains both the verbless *na voljo*, 'at [someone's] disposal' as well as *imeti na voljo*, 'to have at [someone's] disposal', *dati na voljo*, 'to put at [someone's] disposal', *biti na voljo*, 'to be at [someone's] disposal'. The relatively low number of verbs that can be used with *na voljo* seemed to prompt most, but not all of the annotators to include the verb, while others left it out.

Prepositions were frequently contested when in combination with a nominal phrase, e.g. ADJ NOUN - ADP NOUN (*v smislu* 'in [the] sense' vs. *formalnem smislu*, 'formal sense'; *v letih*, 'in [the] years' vs. *zadnjih letih*, 'last years') or ADJ NOUN - ADP ADJ NOUN (*[na] delovnem mestu* '[in] the workplace', *[v] zrelih letih*, lit. 'in mature years', 'at an older age'). Annotators were instructed to consult Slovene corpora to determine the most frequent scope of annotation, but while some interpreted the preposition as an obligatory element, others left it out based on their interpretation, e.g. whether the adjective in the MWE can be considered an open slot (*v [zadnjih/prejšnjih/naslednjih] letih*, 'in the [last/previous/next] years'; similar to numerals in ADP NOUN - ADP NUM NOUN: *pred [desetimi] leti*, '[ten] years ago'; or determiners in ADP DET NOUN - ADP NOUN: *čez [nekaj] dni*, 'in [a few] days') or whether the nominal phrase occurs frequently enough by itself (*delovno mesto*, 'workplace').

There is also some disagreement with regard to the inclusion of auxiliary verbs in verbal MWEs, e.g. AUX VERB ADP NOUN - VERB ADP NOUN (*[je] vzel pod drobnogled*, '[did] take under scrutiny') and AUX VERB NOUN - VERB NOUN (*[ni] odprl ust*, lit. '[didn't] open [his] mouth', 'remained silent'), particularly when there is a negation, but both the negated and non-negated versions are viable (*je odprl usta*, 'he spoke', *ni odprl ust*, 'he remained silent').

4.3. Overlap with Named Entities

Because the SUK 1.0 corpus was also independently annotated with named entities, we analyzed our MWE annotations in terms of tokens that have been annotated as named entities in order to explore any potential legitimate overlaps. Only 334 (3.77%) candidates contain at least one token that has also been annotated as a named entity, and only 115 were annotated by multiple annotators. By analyzing the distribution of the named entity annotations within these 115 candidates, we see that the majority were annotated as organizations (48%) or have no annotation (39%; meaning that not all the MWE elements overlap with the named entity), while other NE categories account for much smaller percentages: miscellaneous (10%), location (2%), person (1%), and person-derivative (0.5%). The guidelines mention that generic titles of institutions, documents, etc. should be annotated as MWEs, particularly if they indicate culturally specific expressions with no direct equivalents or transparent translations in other languages.

A closer look at the examples shows that in the majority of cases, the MWE annotations are nested within NE annotations (e.g. *[Ustavno sodišče] Slovenije*, 'the [Constitutional Court] of Slovenia'; *Urad za [narodnostne manjšine]*, 'Office of [National Minorities]'), but the opposite also occurs, with NEs included in MWEs (*na sončni strani [Alp]*, lit. 'on the sunny side of [the Alps], 'in Slovenia'; *kdor gre na [Dunaj], naj pusti trebuh zunaj*, lit. 'whoever goes to Vienna should leave their stomach outside', 'Vienna is very expensive' or 'large cities are very expensive') or appearing in open slots of MWEs (*so voda na [Lutov] mlin*, lit. 'they are water to [Lut's] mill', 'they provide an advantage to him'). These examples are useful to include in the improved guidelines to exemplify the interplay between MWEs and NEs and to provide clearer instructions on how to annotate mixed candidates.

5. Conclusion

In the paper, we presented the results of the first step of the process of comprehensive MWE annotation in the SUK 1.0 corpus, and conducted a number of quantitative analyses to pinpoint potential weak points in the first version of our annotation guidelines. In particular, the process shows that more instructions and examples are required on how to differentiate between terminological candidates and collocations on one hand, and MWEs on the other. Although the annotators seem to achieve a considerable degree of overlap in terms of annotation scope, for some structures, the scope should be more precisely defined (e.g. the inclusion of auxiliary verbs and closed-class parts-of-speech such as prepositions). In addition, closed-class part-of-speech structures can be pre-extracted in order to generate a list of valid candidates as a reference point for annotators and, potentially, for pre-annotating some of the more trivial syntactic combinations. Pre-annotation with a list of all other MWE-candidates is also an option, but might be more difficult to implement for Slovene, which features a flexible word order and is a morphologically rich language.

Although there has not been much overlap between MWEs and NEs in the annotated examples, the ones that do occur nevertheless show the need for more specific guidelines on when to treat candidates as named entities and how to treat borderline examples (e.g. when the lack of capitalization makes it unclear whether the span denotes a named entity or a generic concept) and mixed candidates (nested MWEs within NEs or vice versa).

In our future work, we intend to use the UniDive MWE annotation guidelines to perform a second step annotation of the identified MWE candidates and determine their categories so that they can be added to the SUK 1.0 corpus alongside their PARSEME verbal MWE equivalents. Once the final annotations have been added to the corpus, a second analysis of outlying examples (either those left unannotated by the majority of annotators or those consistently annotated but not considered MWEs in the final version) can provide additional insight for further MWE identification. In addition, the annotated POS-structures can potentially be compared to the total frequencies of POS-structures within the corpus in order to pinpoint whether certain structures are more typical of MWEs in Slovene in general. Additional statistical analyses on MWE patterns can also be performed by taking into account other annotation layers present in the corpus, such as semantic role labeling and UD dependency relations.

6. Acknowledgements

The study presented in this paper was conducted within the *New Grammar of Modern Standard Slovene: Resource and Methods project* (J6-8256), which was financially supported by the Slovenian Research and Innovation Agency (ARIS) between 2017 and 2020. The authors also acknowledge the financial support from the Slovenian Research and Innovation Agency (research core funding No. P6-0411 - *Language Resources and Technologies for Slovene* and No. P6-0215 - *Slovene Language – Basic, Contrastive, and Applied Studies*).

The authors would like to thank the anonymous reviewers for their valuable insight, and all the annotators who participated in the project: Anna Maria Grego, Tjaša Šoltes, Tajda Liplin Šerbetar, Pia Rednak, Jana Vaupotič, Zala Vidic, Karolina Zgaga, and Kaja Gantar.

7. Ethical Considerations and Limitations

It should be noted that 80% of the people who performed the annotation were university-level students of linguistics, and while they were familiarized with the guidelines and their performance was tested and compared to the performance of experts and considered to be satisfactory in the majority of cases, the annotations need to be interpreted with their background in mind.

In addition, the SUK 1.0 corpus mostly contains written standard Slovene, so the results cannot necessarily be extrapolated to e.g. spoken or non-standard Slovene.

8. Bibliographical References

Kubra Adalı, Tutkum Dinc, Memduh Gokırmak, and Gülşen Eryiğit. 2016. Comprehensive annotation of multiword expressions in turkish. *TurCLing 2016, The First International Conference on Turkic Computational Linguistics at CICLING 2016*, pages 60–66.

Špela Arhar Holdt, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, Luka Terčon, Tina Munda, Slavko Žitnik, Nejc Robida, Neli Blagus, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Taja Kuzman, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2022. Training corpus SUK 1.0. Slovenian language resource repository CLARIN.SI.

Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Simon Krek, Tina Munda, Nejc Robida, Luka Terčon, and Slavko Žitnik. 2023. Nadgradnja učnega korpusa ssj550k v suk 1.0. *Razvoj slovenščine v digitalnem okolju*, pages 119–156.

B. T. Sue Atkins and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford Univer-sity Press, New York.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2020. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*.

Polona Gantar, Jaka Čibej, and Mija Bon. 2019. Slovene multi-word units: Identification, categorization, and representation. In *Computational and Corpus-Based Phraseology*, pages 99–112, Cham. Springer International Publishing.

Polona Gantar and Simon Krek. 2011. Slovene lexical database. *Natural Language Processing, Multilinguality: Sixth International Conference*, pages 1–13.

Polona Gantar, Simon Krek, and Taja Kuzman. 2017. Verbal multiword expressions in slovene. *International Conference on Computational and Corpus-Based Phraseology*, pages 1–13.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Akihiko Kato, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Construction of Large-scale English Verbal Multiword Expression Annotated Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Iztok Kosem, Simon Krek, and Polona Gantar. 2021. Semantic data should no longer exist in isolation: the digital dictionary database of slovenian. *Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion*, pages 81–83.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2018. Training corpus ssj500k 2.1. Slovenian language resource repository CLARIN.SI.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña Ruiz, José Luis Sancho Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, and Tina Munda. 2021. Designing the elexis parallel sense-annotated dataset in 10 european languages. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*, pages 377–395.

Najet Hadj Mohamed, Cherifa Ben Khelil, Agata Savary, Iskandar Keskes, Jean-Yves Antoine, and Lamia Belguith Hadrich. 2022. Annotating verbal multiword expressions in arabic: Assessing the validity of a multilingual annotation procedure. *13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1839–1848.

Johanna Monti, Federico Sangati, and Mihael Arčan. 2015. Ted-mwe: a bilingual parallel corpus with mwe annotation: Towards a methodology for annotating mwes in parallel multilingual corpora. *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 193–197.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati Sangati, Ivelina Stoyanova Stoyanova, and Veronika Vincze. 2018. Parseme multilingual corpus of verbal multiword expressions. *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. Parseme meets universal dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tadej Škvorc, Polona Gantar, and Marko Robnik Šikonja. 2022. Mice: mining idioms with contextual embeddings. *Knowledge-based systems Jan. 2022, vol. 235*, pages 1–11.

Elvis Souza and Claudia Freitas. 2023. Annotation of fixed multiword expressions (MWEs) in a Portuguese Universal Dependencies (UD) treebank: Gathering candidates from three different sources. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 442–450, Belo Horizonte, Brazil. Association for Computational Linguistics.

Light Verb Constructions in Universal Dependencies for South Asian Languages

Abishek Stephen, Daniel Zeman
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800 Praha, Czechia
{stephen, zeman}@ufal.mff.cuni.cz

Abstract

We conduct a morphosyntactic investigation into the light verb constructions (LVCs) or the verbo-nominal predicates in South Asian languages. This work spans the Indo-Aryan and Dravidian language families in treebanks based on Universal Dependencies (UD). For the selected languages we show how well the existing annotation guidelines fare for the LVCs. We also reiterate the importance of the core and oblique distinction in UD and its usefulness for making accurate morphosyntactic annotation judgments for such predicates.

Keywords: light verbs, universal dependencies, multiword expressions

1. Introduction

Universal Dependencies (UD) (de Marneffe et al., 2021) presents a morphosyntactically oriented approach to perform linguistic annotations anchored on binary dependency relations between intra-sentential units. These dependency relations hold primarily between content words, while function words are seen as carriers of morphosyntactic features, which typically "belong" to a content word. Such a mechanism is followed in UD to increase the typological parallelism between languages.[1] The selection of the dependency head gets a little complicated in the case of a multiword expression (MWE) where two or more words combine into a single lexical unit with or without morphosyntactic implications (Masini, 2019). One of the MWE classes where this can be witnessed is the light verb construction (LVC).

LVCs (Section 3) have a peculiar semantic composition that may provoke specific approaches to their syntactic analysis; however, in the case of South Asian languages, profound morphosyntactic clues are available and should be taken into account. The current annotations in the treebanks of these languages in UD treat the LVCs[2] as combinations of lexemes that morphosyntactically behave as single words and mark them using the dependency relation `compound`,[3] or its subtype `compound:lvc`. In the case of South Asian languages this is problematic given the surface-identical noun incorporations and object-verb sequences. We illustrate it on two examples from the treebanks of Hindi (Figures 1 and 2) and Telugu (Figures 3 and 4). In each pair, the first example has an LVC annotated as `compound` while the second example with a similar construction treats the noun as an object (`obj`) of the verb. Our main research question is whether these distinctions are well-motivated and clearly defined based on morphosyntax. It implies some broader questions about argument selection criteria and core vs. oblique distinction in South Asian languages.

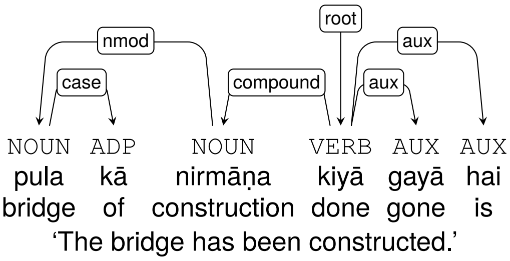

nmod
case
compound
root
aux
aux
NOUN ADP NOUN VERB AUX AUX
pula kā nirmāṇa kiyā gayā hai
bridge of construction done gone is
'The bridge has been constructed.'

Figure 1: A verbo-nominal construction in Hindi (HDTB) annotated as compound.

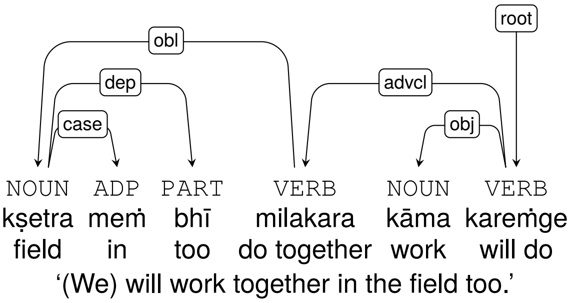

obl
dep
case
root
advcl
obj
NOUN ADP PART VERB NOUN VERB
kṣetra meṁ bhī milakara kāma kareṁge
field in too do together work will do
'(We) will work together in the field too.'

Figure 2: A verbo-nominal construction in Hindi (HDTB) annotated as object.

[1] https://universaldependencies.org/u/overview/syntax.html

[2] For our study we consider all the noun-verb sequences marked as `compound` or `compound:lvc` in the treebanks as LVCs or verbo-nominal predicates.

[3] https://universaldependencies.org/u/dep/compound.html

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 163–177. May 25, 2024.

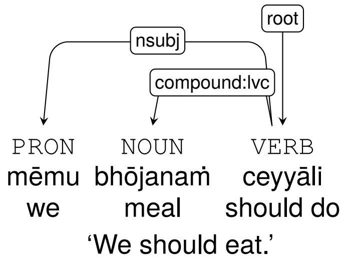


Figure 3: A verbo-nominal construction in Telugu (MTG) annotated as compound.

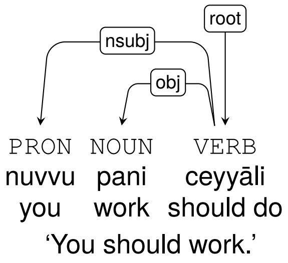


Figure 4: A verbo-nominal construction in Telugu (MTG) annotated as object.

Language	Treebank	Sentences	Words
Sanskrit	Vedic	3,997	27,117
Sanskrit	UFAL	230	1,843
Hindi	HDTB	16,649	351,704
Hindi	PUD	1,000	23,829
Urdu	UDTB	5,130	138,077
Kangri	KDTB	288	2,514
Bhojpuri	BHTB	357	6,665
Bengali	BRU	56	320
Marathi	UFAL	466	3,847
Sinhala	STB	100	880
Telugu	MTG	1,328	6,465
Tamil	TTB	600	9,581
Tamil	MWTT	534	2,584
Malayalam	UFAL	218	2,403

Table 1: Treebank sizes in UD 2.13.

Hence, using the treebanks of Indo-Aryan and Dravidian languages (Table 1) from UD 2.13 (Zeman et al., 2023),[4] we intend to bring to light the fundamental issues around the treatment of various noun-verb sequences. We illustrate that not all noun-verb sequences qualify to be marked as `compound` or `compound:lvc`. We will focus on how the morphosyntactic implications have been overlooked by illustrating supporting examples for the same. Furthermore, we also emphasize the essential distinction between core and oblique arguments in UD (Zeman, 2017) that encompass a crucial role in the morphosyntactic treatment of the noun-verb sequences.

The paper is organized into 6 sections. Discussion of related works happens in Section 2. In Section 3, we present a portrait of LVCs in the selected UD treebanks, organized by language families. In Section 4, we discuss the structural composition of the LVCs by differentiating between incorporation and compounding. In Section 5, the morphosyntax of LVCs finds adequate theoretical treatment, confronted with treebank practice in Section 6.

2. Related Work

Kahane et al. (2018) discusses how to analyze multiword expressions in treebanks based on UD. They mainly focus on distinguishing syntactically irregular MWEs from semantically non-compositional ones and highlight issues related to intra-treebank annotation inconsistencies created because of the MWEs. The analysis concerns the English and French treebanks in UD 2.1 and they note inter-corpus variation in the usage of the dependency relation `compound`. But the LVCs did not receive any attention.

Nivre and Vincze (2015) portrays how LVCs pose interesting challenges for linguistic annotation, especially from a cross-linguistic perspective. They present a survey of the different ways in which LVCs are analyzed in UD 1.1. They group the languages into 3 groups and compare how the LVCs consisting of a transitive verb and a direct object are handled. For example, they report that in the English phrase *take a photo*, *photo* is attached to the verb *take* as a direct object (`dobj`) because the English treebanks in version 1.1 did not distinguish LVCs whereas the treebanks of Swedish, German, and Irish distinguish LVCs through their syntactic structure.

Since our study takes into consideration the constructions labeled as `compound` or `compound:lvc` it is worthwhile to mention that in the Persian treebank (Seraji et al., 2016) the non-canonical subjects are analyzed with respect to LVCs and such constructions are labelled as `compound:lvc`. In the case of the Hungarian treebank (Vincze et al., 2017), the label `dobj:lvc` can be found between the nominal and verbal component of the LVCs, where the `dobj` part of the label marks that syntactically it is a verb–object relation but semantically, it is an LVC, marked by the `lvc` subtype.[5]

Among the South Asian languages, Hindi has received a considerable spotlight for LVCs. Palmer et al. (2009) talks about the LVCs as support-verb

[4] Our analysis will largely be centered around the languages with larger treebanks.

[5] Under UD v2 guidelines this relation is renamed to `obj:lvc`. Besides Hungarian, it is now used also in French and Naija.

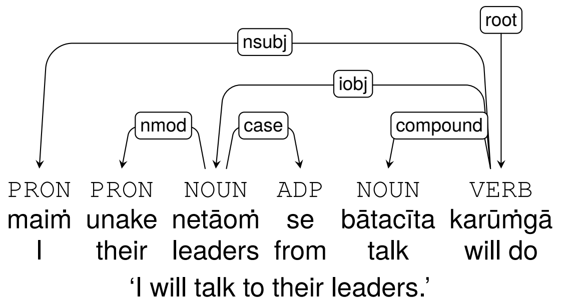

root
nsubj
iobj
nmod
case
compound
PRON PRON NOUN ADP NOUN VERB
maiṁ unake netāoṁ se bātacīta karūṁgā
I their leaders from talk will do
'I will talk to their leaders.'

Figure 5: Compound analysis in Hindi (HDTB).

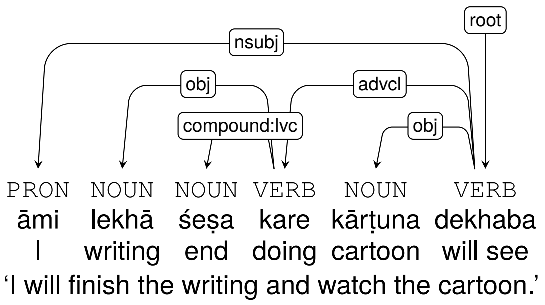

root
nsubj
obj
advcl
compound:lvc
obj
PRON NOUN NOUN VERB NOUN VERB
āmi lekhā śeṣa kare kārṭuna dekhaba
I writing end doing cartoon will see
'I will finish the writing and watch the cartoon.'

Figure 6: Compound analysis in Bengali (BRU).

constructions in Hindi-Urdu where eventive noun phrases combine with several verbs and are analyzed based on case marking. The analysis relies on the Proposition Bank (Palmer et al., 2005) scheme. Begum et al. (2011) focus on the identification of the noun-verb combinations based on the Hindi Dependency Treebank (HDTB).[6] Müller (2019) shows an HPSG analysis and Vaidya et al. (2014) present a TAG (Joshi, 2005) analysis for predicates with the light verbs *karanā* 'to do' and *honā* 'to be' in Hindi, demonstrating that LVCs are a highly productive predicational strategy, challenging for computational grammars.

The PARSEME (Savary et al., 2023) multilingual annotated corpus of verbal multiword expressions also includes Hindi.[7] The underlying hypothesis for the annotations is that verbal MWEs have some degree of semantic non-compositionality and the verb is considered to be the syntactic head.

Within the UD framework, typological studies around LVCs have not involved any of the South Asian languages so far.

3. Light Verb Constructions in UD

The LVCs belong to the class of complex predicates with a wide range of combinatorial potential where a verb (VERB) can combine with adjectives (ADJ), adverbs (ADV) or nouns (NOUN). Out of these, we focus on the verbo-nominal predicates comprising words with the part-of-speech tags NOUN and VERB. This subgroup is most similar to (and confusable with) object-verb sequences; it also has interesting morphosyntactic properties.

3.1. Indo-Aryan Languages

The Indo-Aryan languages are characterized by split ergativity, subject-object agreement, canonical SOV word order, and the presence of postnominal case marking. UD annotation guidelines capture these morphosyntactic nuances aptly although certain inconsistencies remain especially in the case of LVCs. Currently, in UD 2.13, treebanks of Bengali, Bhojpuri, Hindi, Kangri, Marathi, Sanskrit, Sinhala, and Urdu are valid and publicly available. Most of these treebanks use the dependency label `compound` to mark the verbo-nominal compounds or LVCs but the Bengali, Marathi, and Sinhala treebanks use the language-specific dependency sub-type label `compound:lvc`. Figure 5 illustrates a verbo-nominal compound in Hindi *bātacīta karanā* 'to talk' where the verb *karanā* 'to do' selects the noun *bātacīta* 'chit-chat' as the dependent. Other verbs constituting such constructions in the Hindi HDTB and Hindi PUD treebanks include *honā* 'to be', which is the second most frequent verb constituting verbo-nominal predicates after *karanā* 'to do', followed by *lagānā* 'to put'. In Urdu, *denā* 'to give' and *lenā* 'to take' also head verbo-nominal compounds along with *krnā* and *honā*. In Marathi, verbo-nominal compounds function as semantic verbs with varying degrees of lexicalization (Ravishankar, 2017). Here, too, the verbs *karaṇe* 'to do' and *hoṇe* 'to be' are the most frequently selected verbal heads in LVCs. Bengali (Figure 6), Bhojpuri and Kangri also present a similar picture where the verbs 'to do' and 'to be' persistently head such constructions. There are two verbs that function as light verbs in Sinhala, viz. *kara* 'to do', the volitive indicator, and *ve* 'to be', the involitive indicator (Liyanage et al., 2023). The current version of the Sinhala treebank (STB) contains 39 instances of noun-verb combinations marked as `compound:lvc`. Sinhala happens to be the only Indo-Aryan language in UD to select the noun as a head for LVCs (Figure 7).

In the Vedic Sanskrit treebank, complex syntactic structures are expressed through compounds, hence compounds are annotated as if their elements occurred in a non-composed form (Hellwig et al., 2020). Recombination of certain compounds into single words is reported in the Sanskrit UFAL treebank (Dwivedi and Zeman, 2018); the

[6] https://ltrc.iiit.ac.in/treebank_H2014/

[7] https://gitlab.com/parseme/parseme_corpus_hi

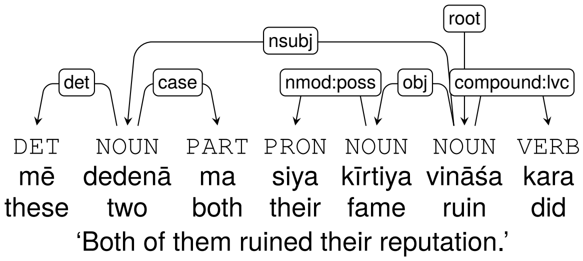

root
nsubj
det
case
nmod:poss
obj
compound:lvc
DET NOUN PART PRON NOUN NOUN VERB
mē dedenā ma siya kīrtiya vināśa kara
these two both their fame ruin did
'Both of them ruined their reputation.'

Figure 7: A verbo-nominal compound in Sinhala (STB), headed by the nominal node.

`compound` relation is not used there.[8] Therefore, we do not find any instance of a verbo-nominal predicate in the current Sanskrit treebanks.

3.2. Dravidian Languages

Within UD, the agglutinating morphology of the Dravidian languages creates multiword tokens (MWTs) or concatenated multiple syntactic words that need to be split during annotation. For example, in Malayalam the copula, complementizer, coordinating clitics, and also occasionally the object and the verb in a sentence occur as a multiword token (Stephen and Zeman, 2023). Similarly, in the Tamil MWTT treebank, the coordinating clitics and the complementizer are split as they are orthographically fused in an MWT. The close resemblance between an MWT and an MWE presents a challenge in the case of the Dravidian languages but morphosyntactic cues come in handy in the disambiguation process. For LVCs, only the compounds with the do-verb *ceyyuka* are labeled as `compound:lvc` in the Malayalam UFAL treebank (Figure 8). The role of a light verb as a verbal licenser is particularly visible in loanwords, which, instead of acquiring the host language verbal morphology, combine with a light verb. An example is Malayalam *aṟasṟṟu ceyyuka* (lit. *to do arrest*) 'to arrest'.

In Tamil MWTT, the noun-verb sequences with the existential be-verb *iru* are marked as `compound:lvc` and the noun is treated as the head selecting the light verb as its dependent (Krishnamurthy and Sarveswaran, 2021), unlike in the Indo-Aryan treebanks. But in the Telugu MTG treebank, the verb is treated as the syntactic head and the noun is considered as the bearer of the predicate semantics for noun-verb sequences marked `compound:lvc` (Rama and Vajjala, 2018). Our overall observation about the Dravidian treebanks is that the distinction between LVCs and regular structures has largely relied on semantic cues or direct influence of the strategy used in the English UD treebanks. Intra-language morphosyntactic clues do not seem to have been considered.

4. Structural Composition of LVCs

According to Butt (2003), the "light" in LVCs indicates that although these constructions respect the standard verb complement schema, the verb cannot be said to be predicating fully but seems to be more of a verbal licenser for nouns. Moreover, the light verbs tend to have a "funny" syntax which distinguishes them from auxiliaries and main verbs. Additionally, Butt (2003) claims that such structures are monoclausal in nature where the predicational elements "co-predicate". Such a view does not align well with saying that they form one lexical (and syntactic) unit, but using the `compound` relation in UD can be understood as saying exactly that. There seems to be a perturbing dichotomy around the lexicality of such sequences as shown in Figure 9, where two instances are analyzed as compounds and one is not. In order to establish a principled position on the structural composition of LVCs, we will now delve into the process of compounding and incorporation and discuss their entanglement with the predicate structure.

4.1. Compounding

We adopt the definition of compounds based on Haspelmath (2023b) as a construction consisting of two strictly adjacent slots for roots[9] that cannot be expanded by full nominal, adjectival, or degree modifiers. Finkbeiner and Schlücker (2019) illustrate the non-expandability on a German example, where the adverb *sehr* 'very' cannot modify the first element in *Alt-bau* 'old building', i.e., **sehr Alt-bau* 'very old building' is not plausible.

On applying Haspelmath's definition to Figure 9, we observe that the noun part of the compound *śurū kara* 'to start' is a root morph whereas the other nouns *golī* 'bullet' and *cunautī* 'challenge' are derived nominal forms of their respective root morphs. If we assume this inference to be accurate, then *cunautī denā* 'to challenge' and *golī calānā* 'to shoot' should not be marked as `compound`. Hence if a noun-verb sequence shall be considered a compound, the nominal part should be a root without suffixes.

[8] Sanskrit UFAL uses the feature Compound=Yes to mark words that were non-final stems within a surface "compound"; however, such forms are treated as separate syntactic words only if the dependency relations between them are other than `compound`.

[9] A root is a contentful morph (i.e., a morph denoting an action, an object, or a property) that can occur as part of a free form without another contentful morph (Haspelmath, 2023b).

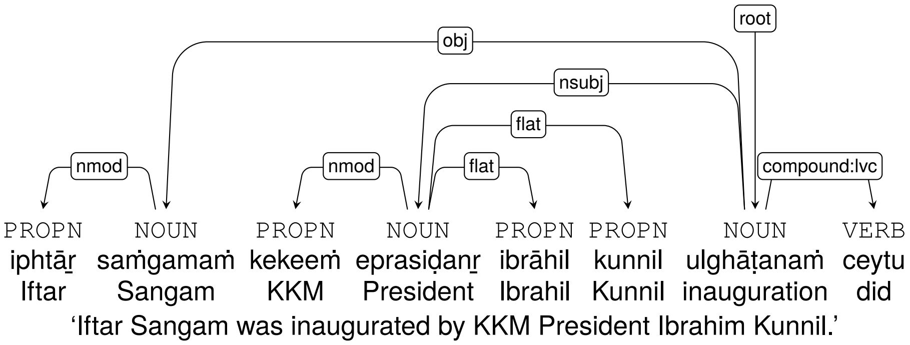

root
obj
nsubj
flat
nmod
nmod
flat
compound:lvc
PROPN NOUN PROPN NOUN PROPN PROPN NOUN VERB
iphtāṟ saṁgamaṁ kekeeṁ eprasiḍanṟ ibrāhil kunnil ulghāṭanaṁ ceytu
Iftar Sangam KKM President Ibrahil Kunnil inauguration did
‘Iftar Sangam was inaugurated by KKM President Ibrahim Kunnil.’

Figure 8: A verbo-nominal compound in Malayalam (UFAL), headed by the nominal node.

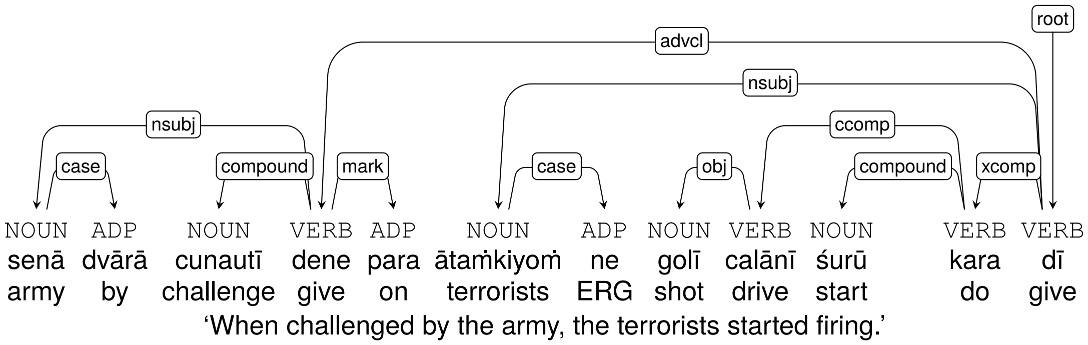

root
advcl
nsubj
nsubj
ccomp
case
compound
mark
case
obj
compound
xcomp
NOUN ADP NOUN VERB ADP NOUN ADP NOUN VERB NOUN VERB VERB
senā dvārā cunautī dene para ātaṁkiyoṁ ne golī calānī śurū kara dī
army by challenge give on terrorists ERG shot drive start do give
‘When challenged by the army, the terrorists started firing.’

Figure 9: Two verbo-nominal compounds: *cunautī denā* ‘to challenge’ and *śurū karanā* ‘to start’. On the other hand, *golī calānā* ‘to shoot’ is annotated just as a verb-object pair (Hindi HDTB).

The UD taxonomy has a more relaxed definition of compounds: it states that the compound relation should be used for combinations of lexemes that morphosyntactically behave as single words, and lexicalization or semantic idiomaticity should not be a criterion for identifying compounds. This entails that a lexicalized expression like *make a decision* in English does not qualify as an MWE or a compound in UD. Expressions that would qualify should have a single argument structure or in other words, the syntactic head of an LVC should select all the required arguments and the dependent noun should neither be modified nor have an argument structure of its own. But in the case of the Indo-Aryan languages, this does not seem to be the case.

In Marathi (Figure 10) the LVC *prayatna karata* ‘trying’ is tagged as compound:lvc where the noun *prayatna* ‘try’ heads the nsubj and xcomp dependency relations which is not consistent with the UD guidelines. For once we could assume it to be a language-specific decision but there are also examples like Figure 11 which say otherwise. In both the examples (Figure 10 and 11) the compound:lvc relation is headed by the verb *karaṇe* ‘to do’ but the dependent nouns are different. This leads a UD user to the conclusion that in such predicates the nouns have arbitrarily chosen argument structure as no morphosyntactic motivations can be seen in the surface syntactic structure. Similar inconsistencies can also be found in other Indo-Aryan languages. This inconsistent behavior suggests that the annotation choices made for the LVCs are not strongly based on a concrete morphosyntactic mechanism.

Among Dravidian languages, Tamil and Malayalam have taken a left-headed approach considering the noun as the head whereas Telugu treats the verb as the syntactic head making the compound:lvc relation right-headed. The annotation of the LVCs is comparatively more consistent than in the Indo-Aryan languages but it seems to be heavily influenced by semantics or by the treatment of LVCs in the English treebanks. For example, the current version of the Malayalam UFAL treebank uses the compound:lvc relation for noun-verb and verb-verb sequences where the do-verb *ceyyuka* appears. No morphosyntactic motivation can be found in the respective documentation pages of the Dravidian languages.

We conclude that if a noun-verb construction is marked as compound(:lvc), the syntactic head is eligible for modifications but not the dependent. If we need to annotate a child of the dependent node in the noun-verb sequence, then the sequence should be treated as verb with object.

4.2. Noun Incorporation

It is also worthwhile to mention the broader typological definition of incorporation by Haspelmath (2023a) according to which an incorporation is an event-denoting noun-verb compound construction in which the noun occupies an argument slot of the verb and occurs in a position where nominal patient arguments cannot occur. In most Indo-Aryan languages, verbo-nominal predicates must be analyzed as a lexical category but paradoxically enough, the noun is on par with a syntactically independent argument (Mohanan, 1995). Therefore, even though noun incorporation is a type of compounding of a syntactic object with the verb, both the object and the verb can have their own argument structures. It may thus be hard to find incorporation that satisfies Haspelmath's definition in South Asian languages. Currently, the UD taxonomy has no special provisions to define incorporation and they are treated as compounds. As a result, there are no distinct annotations for an object-verb pair and a 'conjunct verb'.[10] The Hindi HDTB treebank in UD is converted from the Paninian Dependencies and in that scheme, conjunct verbs have a special tag `pof` (Tandon et al., 2016). It does not denote a dependency but rather represents the fact that the noun-verb sequence is an MWE. The logic behind the usage of the `pof` tag is based on the semantic coherence of the noun-verb sequence being a single predicative element although some morphosyntactic cues do come in handy (discussed in Section 5). Tandon et al. (2016) also acknowledges that the identification of conjunct verbs is problematic as it appears to be an issue for the syntax-semantics interface and the decision was left to the annotators at the cost of inconsistencies in the data. On conversion from the Paninian dependencies to UD all the `pof` relations were automatically changed to `compound` and the inconsistencies persist. This brings us to a juncture where distinguishing object-verb sequences from noun incorporation becomes necessary. For Dravidian languages, Sudharsan (1998) states that if the noun in a noun-verb sequence cannot be inflected for case or number and even cannot be modified by an adjective then it is the case of a noun incorporated into the verb. Since incorporated nouns do not take case or plural markers and external modifiers, they are morphosyntactically different from the regular object nouns. Similarly for Indo-Aryan languages or more specifically for Hindi-Urdu, Mohanan (2017) has also recommended very similar criteria for distinguishing objects and incorporated nouns. These criteria treat noun incorporation as a type of compounding but there are also cases where such syntactic tests are inadequate, for example in cases of independent syntactic argument structures. The nominal part can be a noun or a root morph. Usually, the root morphs do not have an argument structure of their own but a noun on the other hand has the potential to have its own argument structure in such noun-verb constructions (Mohanan, 1995). To qualify for a `compound:lvc` relation the noun-verb sequence should have a single argument structure but that is not always true in case of noun incorporations. This indicates a need for a distinction between compounding and noun incorporation. In the following section, we find taxonomical differences between them but it will be also worthwhile to test how similar their morphosyntax is and how we can distinguish them from object-verb sequences.

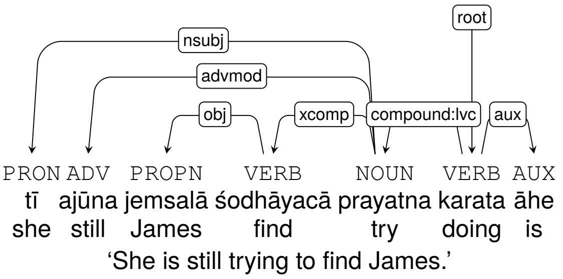

root
nsubj
advmod
obj
xcomp
compound:lvc
aux
PRON ADV PROPN VERB NOUN VERB AUX
tī ajūna jemsalā śodhāyacā prayatna karata āhe
she still James find try doing is
'She is still trying to find James.'

Figure 10: A verbo-nominal compound in Marathi (UFAL), arguments attached to the nominal node.

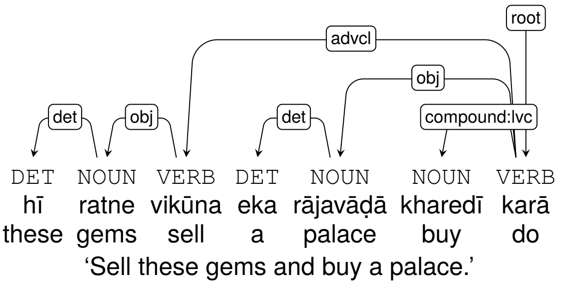

root
advcl
obj
det
obj
det
compound:lvc
DET NOUN VERB DET NOUN NOUN VERB
hī ratne vikūna eka rājavāḍā kharedī karā
these gems sell a palace buy do
'Sell these gems and buy a palace.'

Figure 11: A verbo-nominal compound in Marathi (UFAL), arguments attached to the verbal node.

5. Morphosyntax of LVCs

Subjects and objects in UD must satisfy the condition of being core arguments, which means that they should receive the language-specific coding and treatment associated with the grammatical functions **S**, **A**, and **P** (Zeman, 2017; Andrews, 2007). This coding derives from primary transitive predicates and may include various strategies,

[10] Conjunct verb is a term often used by Indian linguists. In complex predicates, Noun/Adjective-Verb combinations are called 'conjunct verbs' and Verb-Verb combinations are called 'compound verbs' (Begum et al., 2011). But as stated earlier, we define compounds differently based on UD taxonomy.

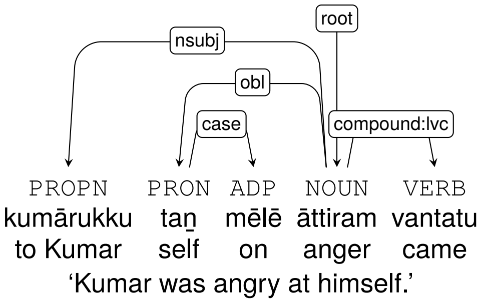

root
nsubj
obl
case
compound:lvc
PROPN PRON ADP NOUN VERB
kumārukku taṉ mēlē āttiram vantatu
to Kumar self on anger came
‘Kumar was angry at himself.’

Figure 12: A verbo-nominal compound in Tamil (MWTT), headed by the nominal node.

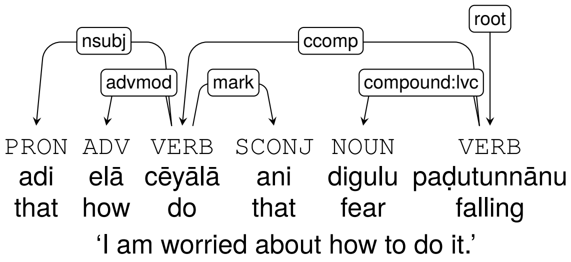

root
nsubj
ccomp
advmod
mark
compound:lvc
PRON ADV VERB SCONJ NOUN VERB
adi elā cēyālā ani digulu paḍutunnānu
that how do that fear falling
‘I am worried about how to do it.’

Figure 13: A verbo-nominal compound in Telugu (MTG).

including case marking on nouns and agreement morphology on verbs. Nominals whose grammatical function is **A** or **S** are called subjects and their dependency relation to the verb is `nsubj` whereas the nominals whose grammatical function is **P** are called (direct) objects and their dependency relation to the verb is `obj` (Zeman, 2017). Turning back to Haspelmath's definition of noun incorporation in Section 4, the incorporated noun cannot occupy the *patient* position and cannot have the function **P**. Hence, we illustrate the behavior of LVCs through morphosyntactic processes like verbal agreement, case marking, and nominal modification. This analysis will bring out the distinctions between compounds and object-verb sequences.

5.1. Case Marking

Hindi, Urdu, and some other Indo-Aryan languages follow a split-ergative pattern. Perfective clauses have the ergative alignment, imperfective clauses have a nominative-accusative alignment. In the latter, the subject is in the bare nominative form (without adpositions), while animate direct objects use the postposition *ko*. Inanimate direct objects may omit the postposition *ko*; if they use it, the object is understood as definite. The accusative (oblique) case is used with the postposition, but without it, the object stays in nominative. Indirect objects always use the postposition *ko*. In transitive perfective clauses, the subject takes the ergative postposition *ne*.

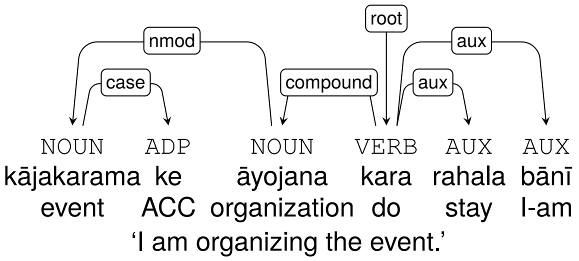

root
nmod
aux
case
compound
aux
NOUN ADP NOUN VERB AUX AUX
kājakarama ke āyojana kara rahala bānī
event ACC organization do stay I-am
‘I am organizing the event.’

Figure 14: A verbo-nominal compound in Bhojpuri (BHTB) where the nominal conjunct *āyojana* ‘organizing’ selects the argument *kājakarama* ‘event’ case marked using the postposition *ke* ‘ACC’.

Nominal parts of LVC candidates are inanimate and thus harder to distinguish from direct objects. However, the ability to take the optional *ko* signals that the noun is an object.

A few true LVCs, such as *śurū karanā* ‘to start’, can be transitive as a whole. Here, *śurū* is not an object and the whole compound may take a real object (which follows the above criteria for objects) or a complement clause. In most cases, however, the nominal part of the LVC is a direct object, and if the whole LVC is semantically transitive, then the external “object” is coded as a nominal modifier (with the genitive postposition *kā*) of the noun in the LVC. It should then be annotated as `nmod` in UD (*pula kā nirmāṇa* ‘construction of bridge’ in Figure 1). Even with *śurū karanā* the genitive strategy is a possible alternative and occurred twice in HDTB. The predicating nominals in Hindi may also select arguments with other postpositions, such as *par* ‘on’, *se* ‘from’, or *ko* ‘to’ (Vaidya et al., 2016).

Eastern Indo-Aryan languages such as Bhojpuri do not have the ergative alignment in perfective clauses. Similarly to Hindi, animacy and definiteness play a role in marking of the direct object (Thakur, 2021). However, Bhojpuri uses the same postposition *(ke)* (Figure 14) for accusative, dative, and genitive, making it less obvious when it is selected by the nominal and not the verb.

In Dravidian languages too the arguments are postpositionally case-marked but in an agglutinative manner. In Tamil MWTT, we find examples like *kumār muṉṉukku vantāṉ* ‘Kumar progressed (in his career/ life)’ where the nominal component *muṉṉukku* ‘to the front’ of the `compound:lvc` is assigned the dative case and the subject proper noun *Kumar* takes the nominative case. Since *muṉṉukku* is treated as the `root` the analysis gets blurry but *muṉṉukku vā* ‘to progress’ might not qualify to be considered as a compound due to the dative case marking.

The presence of an adpositional phrase selected by the nominal differentiates compounding from noun incorporation but this does not provide a suitable distinction between object-verb sequences and noun incorporations at least for the Indo-Aryan languages. In this light, we observe that currently most of the `compound:lvc` or `compound` relations describing noun-verb sequences are not true compounds as the nominal participant does show case marking.

5.2. Agreement

The split-ergative pattern in some Indo-Aryan languages allows for testing of object-verb agreement. In imperfective clauses, the gender and number of the subject are cross-referenced by the verb's morphology. In transitive perfective clauses, the ergative postposition *ne* blocks agreement with the subject; but unless the direct object is marked with *ko*, verbal morphology cross-references the gender and number of the object (rather than subject). If the postposition *ko* is present, the verb takes the default masculine singular form.[11]

Agreement with the verb in transitive-perfective clauses is another signal that the nominal of an LVC candidate is an object rather than part of a compound. And it can also attest to the opposite: In *mere pitā ne pūjā śurū kar dī hai* 'my father has started the prayer', the verb has a feminine form, agreeing with *pūjā*, while both *pitā* 'father' and *śurū* 'start' are masculine.

Eastern Indo-Aryan languages (e.g., Bhojpuri and Bengali), as well as Dravidian languages, follow the nominative-accusative pattern with subject-predicate agreement and no ergativity (Krishnamurti, 2003). In Telugu, the verb agrees with the subject when it is in the nominative case, whereas when there is a dative "subject", the verb agrees with the incorporated noun (Nadimpalli and Lakshmi, 2022). Similar observations can be made for other Dravidian languages except for Malayalam where subject-verb agreement is absent.

To conclude this section, in many instances of noun-verb sequences agreement between the noun and the verb is observed and represents a deviation from typical compound behavior.

5.3. Modification

One of the signs of compounds is that their parts (and especially the dependent part) cannot be modified individually. We have seen that the patient in Hindi LVC candidates is often encoded as a modifier of the predicative nominal, which speaks against a noun-verb compound analysis. Similarly, in Kangri in Figure 15, the nominal *galla* 'matter' is modified by the determiner *isadī* 'this', suggesting that *galla mannī* is not a compound.

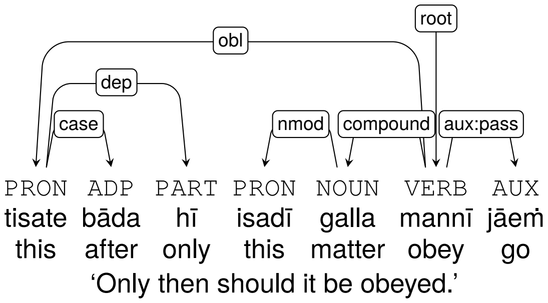

root
obl
dep
case
nmod
compound
aux:pass
PRON ADP PART PRON NOUN VERB AUX
tisate bāda hī isadī galla mannī jāeṁ
this after only this matter obey go
'Only then should it be obeyed.'

Figure 15: Compound analysis in Kangri (KDTB).

In Telugu too, we find similar instances of the predicative nominal modification. For example, in *vāḍu cālā takkuva pani cēsēḍu* 'He does very little work', *takkuva* 'less' modifies *pani* 'work' which happens to be in a `compound:lvc` relation with *cēsēḍu* 'do'.

5.4. Word Order

Real compounds would not allow intervening words between the noun and the verb (at least not by Haspelmath's definition of compounds). An intervention seems to be always possible at least by the negative particle: *unhoṁne batāyā ki abhī pahale baica kā praśikṣaṇa śurū* ***nahīṁ*** *huā hai.* 'He told that the training of the first batch has not started yet.'

5.5. Transitivity

The grammars of Indo-Aryan languages feature a systematic opposition of transitive (causative) and intransitive verbs. The intransitive counterpart of *karanā* in Hindi is *honā* 'to be, become, happen'; as shown in Section 3, its cognates do the same job in the other languages. Whenever it is inappropriate to analyze *X karanā* as a compound, the same can be said about *X honā*. However, as *honā* is intransitive, *X* can hardly act as its object. In Hindi-Urdu this verb is also used as the copula, hence a copular analysis may be an alternative. Where the light verb cannot be a copula, we should probably go with secondary predication (`xcomp`).

6. LVCs in UD Revisited

Noun-verb compounds are very frequent in the current UD treebanks of South Asian languages. In Hindi HDTB, there are 6187 such compounds with the 5 most common verbs alone (out of which 4159 occurrences belong just to *karanā* 'to do'). A similar pattern is found in the smaller Urdu treebank:

[11] While in general postpositions block agreement in Indo-Aryan languages, Gujarati is an exception where verb agreement works despite postpositions (Subbarao, 2012, p. 97).

3542 occurrences with the top 5 verbs, including 2346 with *krnā* 'to do'. The remaining treebanks are an order of magnitude smaller, yet we find 58 different compounds in Bhojpuri and 31 in Hindi PUD occurring twice or more. Nevertheless, the treebanks are not always consistent and it is not uncommon to see the same noun-verb combination annotated sometimes as a compound and sometimes as an object.

For example, Hindi *bāta karanā* 'to talk' is a relatively frequent expression and it is usually annotated as `compound` (118 instances), though occasionally it is annotated as `obj` (25 instances). The noun *bāta* can occur with the postposition *ko* and then it is always annotated as the object (13 instances). It can occur in the plural (11 instances without *ko* and 2 instances with *ko*) and there can occasionally be other constituents between it and the verb. In transitive perfective clauses, the verb agrees with its feminine gender: *Naṭavara Siṁha* (Masc) *ne Nirupama Sena se bāta* (Fem) *kī* (Fem) *hai* 'Natwar Singh had spoken to Nirupam Sen'. The noun *bāta* can be also modified by a nominal denoting the matter that is being talked about. All this is evidence that *bāta* should be syntactically analyzed as the object of *karanā*. For more statistics across the treebanks, see the Appendix.

Furthermore, based on the arguments present in Section 5, we can conclude that in the present versions of the treebanks of South Asian languages, the treatment of noun-verb sequences or LVCs as compounds is not consistent because the interplay of surface level similarities between real noun-verb compounds and noun incorporations somehow weigh down the morphosyntatic cues. There should not be a problem if noun-verb compounds satisfying the UD guidelines are marked as `compound:lvc` just to differentiate it from other type of compounds. This would also handle most of the noun incorporations, but once the nominal participant is case marked, modified or triggering verbal agreement, the sequence should be analyzed differently. One of the solutions could be to label the relation `obj:lvc`, modifying Vincze et al. (2017)'s proposal to fit the current UD version. By doing so, there will be a three-way distinction between noun-verb compounds and noun incorporations (with a single argument structure) marked as `compound:lvc`, object-verb sequences marked as `obj` and noun-incorporations with individual noun and verb argument structures as `obj:lvc`.

7. Conclusion

We have presented morphosyntactic clues for identifying light verb constructions in South Asian languages, which could prove instrumental in achieving consistent annotations of `compound` and `compound:lvc` dependency relations. While LVCs as semantically idiosyncratic constructions are widespread in these languages, we have shown that in many cases their syntactic behavior is transparent or very close to standard object-verb constructions. Their compound analysis should be reconsidered and the annotation could be changed to `obj` or `obj:lvc` based on the type of argument sharing.

We also touched upon the core vs oblique distinctions and highlighted the phenomenon of noun incorporations, which can be beneficial for tackling similar inconsistencies beyond the languages handled in this study.

8. Acknowledgements

This work was supported by the Grant 20-16819X (LUSyD) of the Czech Science Foundation (GAČR); and LM2023062 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic; and the Charles University project GA UK No. 101924; and partially supported by SVV project number 260 698.

9. Bibliographical References

Avery D. Andrews. 2007. The major functions of the noun phrase. In Timothy Shopen, editor, *Language Typology and Syntactic Description*, page 132–223. Cambridge University Press, Cambridge, UK.

Rafiya Begum, Karan Jindal, Ashish Jain, Samar Husain, and Dipti Misra Sharma. 2011. Identification of conjunct verbs in Hindi and its effect on parsing accuracy. In *Computational Linguistics and Intelligent Text Processing - 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part I*, volume 6608 of *Lecture Notes in Computer Science*, pages 29–40. Springer.

Miriam Butt. 2003. The light verb jungle. *Harvard Working Papers in Linguistics*, 9.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Puneet Dwivedi and Daniel Zeman. 2018. The forest lion and the bull: Morphosyntactic annotation of the Panchatantra. *Computación y Sistemas*, 22(4):1377–1384.

Rita Finkbeiner and Barbara Schlücker. 2019. *Compounds and multi-word expressions in the languages of Europe*, pages 1–44. De Gruyter.

Martin Haspelmath. 2023a. Compound and incorporation constructions as combinations of unexpandable roots.

Martin Haspelmath. 2023b. Defining the word. *WORD*, 69(3):283–297.

Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. The treebank of vedic Sanskrit. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5137–5146, Marseille, France. European Language Resources Association.

Aravind K. Joshi. 2005. 483 Tree-Adjoining Grammars. In *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Sylvain Kahane, Kim Gerdes, and Marine Courtin. 2018. Multi-word annotation in syntactic treebanks: Propositions for universal dependencies. In *16th international conference on Treebanks and Linguistic Theories (TLT)*.

Parameswari Krishnamurthy and Kengatharaiyer Sarveswaran. 2021. Towards building a modern written Tamil treebank. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 61–68, Sofia, Bulgaria. Association for Computational Linguistics.

Bhadriraju Krishnamurti. 2003. Syntax. In *The Dravidian Languages*, Cambridge Language Surveys, pages 420–469. Cambridge University Press, Cambridge, UK.

Chamila Liyanage, Kengatharaiyer Sarveswaran, Thilini Nadungodage, and Randil Pushpananda. 2023. Sinhala dependency treebank (STB). In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 17–26, Washington, D.C. Association for Computational Linguistics.

Francesca Masini. 2019. Multi-word expressions and morphology.

Tara Mohanan. 1995. Wordhood and lexicality: Noun incorporation in Hindi. *Natural Language & Linguistic Theory*, 13(1):75–134.

Tara Mohanan. 2017. *Grammatical and Light Verbs*, pages 1–27. John Wiley & Sons, Ltd.

Stefan Müller. 2019. Complex predicates: Structure, potential structure and underspecification. In *Linguistic Issues in Language Technology (LiLT) 16*.

Satish Kumar Nadimpalli and Bh VN Lakshmi. 2022. Is there noun incorporation in Telugu? *Journal of Language and Linguistic Studies*, 18(2):895–903.

Joakim Nivre and Veronika Vincze. 2015. Light verb constructions in universal dependencies. In *Poster at the 5th PARSEME meeting, Iasi, Romania*.

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Taraka Rama and Sowmya Vajjala. 2018. A dependency treebank for Telugu. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 119–128, Prague, Czechia.

Vinit Ravishankar. 2017. A Universal Dependencies treebank for Marathi. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 190–200, Prague, Czechia.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Mojgan Seraji, Filip Ginter, and Joakim Nivre. 2016. Universal dependencies for Persian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2361–2365.

Abishek Stephen and Daniel Zeman. 2023. Universal Dependencies for Malayalam. *The Prague Bulletin of Mathematical Linguistics*, 120:31–46.

Karumuri V. Subbarao. 2012. *South Asian Languages: A Syntactic Typology*. Cambridge University Press, Cambridge, UK.

Anuradha Sudharsan. 1998. *A minimalist account of null subjects in Kannada*. Ph.D. thesis, University of Hyderabad.

Juhi Tandon, Himani Chaudhry, Riyaz Ahmad Bhat, and Dipti Sharma. 2016. Conversion from Paninian karakas to Universal Dependencies for Hindi dependency treebank. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 141–150, Berlin, Germany. Association for Computational Linguistics.

Gopal Thakur. 2021. *A Grammar of Bhojpuri*. LINCOM studies in Indo-European linguistics. LINCOM GmbH.

Ashwini Vaidya, Sumeet Agarwal, and Martha Palmer. 2016. Linguistic features for Hindi light verb construction identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1320–1329.

Ashwini Vaidya, Owen Rambow, and Martha Palmer. 2014. Light verb constructions with 'do' and 'be' in Hindi: A tag analysis. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 127–136.

Veronika Vincze, Katalin Ilona Simkó, Zsolt Szántó, and Richárd Farkas. 2017. Universal Dependencies and morphology for Hungarian – and on the price of universality. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 356–365, Valencia, Spain. Association for Computational Linguistics.

Daniel Zeman. 2017. Core arguments in Universal Dependencies. In *Proceedings of the fourth international conference on dependency linguistics (DepLing 2017)*, pages 287–296, Pisa, Italy.

10. Language Resource References

Zeman, Daniel and Nivre, Joakim and Abrams, Mitchell and Ackermann, Elia and Aepli, Noëmi and Aghaei, Hamid and Agić, Željko and Ahmadi, Amir and Ahrenberg, Lars and Ajede, Chika Kennedy and Akkurt, Salih Furkan and Aleksandravičiūtė, Gabrielė and Alfina, Ika and Algom, Avner and Alnajjar, Khalid and Alzetta, Chiara and Andersen, Erik and Antonsen, Lene and Aoyama, Tatsuya and Aplonova, Katya and Aquino, Angelina and Aragon, Carolina and Aranes, Glyd and Aranzabe, Maria Jesus and Arıcan, Bilge Nas and Arnardóttir, Þórunn and Arutie, Gashaw and Arwidarasti, Jessica Naraiswari and Asahara, Masayuki and Ásgeirsdóttir, Katla and Aslan, Deniz Baran and Asmazoğlu, Cengiz and Ateyah, Luma and Atmaca, Furkan and Attia, Mohammed and Atutxa, Aitziber and Augustinus, Liesbeth and Avelãs, Mariana and Badmaeva, Elena and Balasubramani, Keerthana and Ballesteros, Miguel and Banerjee, Esha and Bank, Sebastian and Barbu Mititelu, Verginica and Barkarson, Starkaður and Basile, Rodolfo and Basmov, Victoria and Batchelor, Colin and Bauer, John and Bedir, Seyyit Talha and Behzad, Shabnam and Belieni, Juan and Bengoetxea, Kepa and Benli, İbrahim and Ben Moshe, Yifat and Berk, Gözde and Bhat, Riyaz Ahmad and Biagetti, Erica and Bick, Eckhard and Bielinskienė, Agnė and Bjarnadóttir, Kristín and Blokland, Rogier and Bobicev, Victoria and Boizou, Loïc and Borges Völker, Emanuel and Börstell, Carl and Bosco, Cristina and Bouma, Gosse and Bowman, Sam and Boyd, Adriane and Braggaar, Anouck and Branco, António and Brokaitė, Kristina and Burchardt, Aljoscha and Campos, Marisa and Candito, Marie and Caron, Bernard and Caron, Gauthier and Carvalheiro, Catarina and Carvalho, Rita and Cassidy, Lauren and Castro, Maria Clara and Castro, Sérgio and Cavalcanti, Tatiana and Cebiroğlu Eryiğit, Gülşen and Cecchini, Flavio Massimiliano and Celano, Giuseppe G. A. and Čéplö, Slavomír and Cesur, Neslihan and Cetin, Savas and Çetinoğlu, Özlem and Chalub, Fabricio and Chamila, Liyanage and Chauhan, Shweta and Chi, Ethan and Chika, Taishi and Cho, Yongseok and Choi, Jinho and Chun, Jayeol and Chung, Juyeon and Cignarella, Alessandra T. and Cinková, Silvie and Collomb, Aurélie and Çöltekin, Çağrı and Connor, Miriam and Corbetta, Claudia and Corbetta, Daniela and Costa, Francisco and Courtin, Marine and Crabbé, Benoît and Cristescu, Mihaela and Cvetkoski, Vladimir and Dale, Ingerid Løyning and Daniel, Philemon and Davidson, Elizabeth and de Alencar, Leonel Figueiredo and Dehouck, Mathieu and de Laurentiis, Martina and de Marneffe, Marie-Catherine and de Paiva, Valeria and Derin, Mehmet Oguz and de Souza, Elvis and Diaz de Ilarraza, Arantza and Dickerson, Carly and Dinakaramani, Arawinda and Di Nuovo, Elisa and Dione, Bamba and Dirix, Peter and Dobrovoljc, Kaja and Doyle, Adrian and

Dozat, Timothy and Droganova, Kira and Duran, Magali Sanches and Dwivedi, Puneet and Ebert, Christian and Eckhoff, Hanne and Eguchi, Masaki and Eiche, Sandra and Eli, Marhaba and Elkahky, Ali and Ephrem, Binyam and Erina, Olga and Erjavec, Tomaž and Essaidi, Farah and Etienne, Aline and Evelyn, Wograine and Facundes, Sidney and Farkas, Richárd and Favero, Federica and Ferdaousi, Jannatul and Fernanda, Marília and Fernandez Alcalde, Hector and Fethi, Amal and Foster, Jennifer and Fransen, Theodorus and Freitas, Cláudia and Fujita, Kazunori and Gajdošová, Katarína and Galbraith, Daniel and Gamba, Federica and Garcia, Marcos and Gärdenfors, Moa and Gerardi, Fabrício Ferraz and Gerdes, Kim and Gessler, Luke and Ginter, Filip and Godoy, Gustavo and Goenaga, Iakes and Gojenola, Koldo and Gökırmak, Memduh and Goldberg, Yoav and Gómez Guinovart, Xavier and González Saavedra, Berta and Griciūtė, Bernadeta and Grioni, Matias and Grobol, Loïc and Grūzītis, Normunds and Guillaume, Bruno and Guiller, Kirian and Guillot-Barbance, Céline and Güngör, Tunga and Habash, Nizar and Hafsteinsson, Hinrik and Hajič, Jan and Hajič jr., Jan and Hämäläinen, Mika and Hà Mỹ, Linh and Han, Na-Rae and Hanifmuti, Muhammad Yudistira and Harada, Takahiro and Hardwick, Sam and Harris, Kim and Haug, Dag and Heinecke, Johannes and Hellwig, Oliver and Hennig, Felix and Hladká, Barbora and Hlaváčová, Jaroslava and Hociung, Florinel and Hohle, Petter and Huang, Yidi and Huerta Mendez, Marivel and Hwang, Jena and Ikeda, Takumi and Ingason, Anton Karl and Ion, Radu and Irimia, Elena and Ishola, Ọlájídé and Islamaj, Artan and Ito, Kaoru and Jagodzińska, Sandra and Jannat, Siratun and Jelínek, Tomáš and Jha, Apoorva and Jiang, Katharine and Johannsen, Anders and Jónsdóttir, Hildur and Jørgensen, Fredrik and Juutinen, Markus and Kaşıkara, Hüner and Kabaeva, Nadezhda and Kahane, Sylvain and Kanayama, Hiroshi and Kanerva, Jenna and Kara, Neslihan and Karahóğa, Ritván and Kåsen, Andre and Kayadelen, Tolga and Kengatharaiyer, Sarveswaran and Kettnerová, Václava and Kharatyan, Lilit and Kirchner, Jesse and Klementieva, Elena and Klyachko, Elena and Kocharov, Petr and Köhn, Arne and Köksal, Abdullatif and Kopacewicz, Kamil and Korkiakangas, Timo and Köse, Mehmet and Koshevoy, Alexey and Kotsyba, Natalia and Kovalevskaitė, Jolanta and Krek, Simon and Krishnamurthy, Parameswari and Kübler, Sandra and Kuqi, Adrian and Kuyrukçu, Oğuzhan and Kuzgun, Aslı and Kwak, Sookyoung and Kyle, Kris and Laan, Käbi and Laippala, Veronika and Lambertino, Lorenzo and Lando, Tatiana and Larasati, Septina Dian and Lavrentiev, Alexei and Lee, John and Lê Hồng, Phương and Lenci, Alessandro and Lertpradit, Saran and Leung, Herman and Levina, Maria and Levine, Lauren and Li, Cheuk Ying and Li, Josie and Li, Keying and Li, Yixuan and Li, Yuan and Lim, KyungTae and Lima Padovani, Bruna and Lin, Yi-Ju Jessica and Lindén, Krister and Liu, Yang Janet and Ljubešić, Nikola and Lobzhanidze, Irina and Loginova, Olga and Lopes, Lucelene and Lusito, Stefano and Luthfi, Andry and Luukko, Mikko and Lyashevskaya, Olga and Lynn, Teresa and Macketanz, Vivien and Mahamdi, Menel and Maillard, Jean and Makarchuk, Ilya and Makazhanov, Aibek and Mandl, Michael and Manning, Christopher and Manurung, Ruli and Marşan, Büşra and Mărănduc, Cătălina and Mareček, David and Marheinecke, Katrin and Markantonatou, Stella and Martínez Alonso, Héctor and Martín Rodríguez, Lorena and Martins, André and Martins, Cláudia and Mašek, Jan and Matsuda, Hiroshi and Matsumoto, Yuji and Mazzei, Alessandro and McDonald, Ryan and McGuinness, Sarah and Mendonça, Gustavo and Merzhevich, Tatiana and Miekka, Niko and Miller, Aaron and Mischenkova, Karina and Missilä, Anna and Mititelu, Cătălin and Mitrofan, Maria and Miyao, Yusuke and Mojiri Foroushani, AmirHossein and Molnár, Judit and Moloodi, Amirsaeid and Montemagni, Simonetta and More, Amir and Moreno Romero, Laura and Moretti, Giovanni and Mori, Shinsuke and Morioka, Tomohiko and Moro, Shigeki and Mortensen, Bjartur and Moskalevskyi, Bohdan and Muischnek, Kadri and Munro, Robert and Murawaki, Yugo and Müürisep, Kaili and Nainwani, Pinkey and Nakhlé, Mariam and Navarro Horñiacek, Juan Ignacio and Nedoluzhko, Anna and Nešpore-Bērzkalne, Gunta and Nevaci, Manuela and Nguyễn Thị, Lương and Nguyễn Thị Minh, Huyền and Nikaido, Yoshihiro and Nikolaev, Vitaly and Nitisaroj, Rattima and Nourian, Alireza and Nunes, Maria das Graças Volpe and Nurmi, Hanna and Ojala, Stina and Ojha, Atul Kr. and Óladóttir, Hulda and Olúòkun, Adédayọ̀ and Omura, Mai and Onwuegbuzia, Emeka and Ordan, Noam and Osenova, Petya and Östling, Robert and Øvrelid, Lilja and Özateş, Şaziye Betül and Özçelik, Merve and Özgür, Arzucan and Öztürk Başaran, Balkız and Paccosi, Teresa and Palmero Aprosio, Alessio and Panova, Anastasia and Pardo, Thiago Alexandre Salgueiro and Park, Hyunji Hayley and Partanen, Niko and Pascual, Elena and Passarotti, Marco and Patejuk, Agnieszka and Paulino-Passos, Guilherme and Pedonese, Giulia and Peljak-Łapińska, Angelika and Peng, Siyao and Peng, Siyao Logan and Pereira, Rita and Pereira, Sílvia and Perez, Cenel-Augusto and Perkova, Natalia and Perrier, Guy and Petrov, Slav and Petrova, Daria and Peverelli, Andrea and Phelan, Jason and Pierre-Louis, Claudel and Piitulainen, Jussi and Pinter, Yuval and Pinto, Clara and Pintucci, Rodrigo and Pirinen, Tommi A and Pitler, Emily and Plamada, Magdalena and Plank, Barbara and Poibeau, Thierry and Ponomareva, Larisa and Popel, Martin and Pretkalniņa, Lauma and Prévost, Sophie and Prokopidis, Prokopis and Przepiórkowski, Adam and Pugh, Robert and Puolakainen, Tiina and Pyysalo, Sampo and Qi, Peng and Querido, Andreia and Rääbis, Andriela and Rademaker, Alexandre and Rahoman, Mizanur and Rama, Taraka and Ramasamy, Loganathan and Ramisch, Carlos and Ramos, Joana and Rashel, Fam and Rasooli, Mohammad Sadegh and Ravishankar, Vinit and Real, Livy and Rebeja, Petru and Reddy, Siva and Regnault, Mathilde and Rehm, Georg and Riabi, Arij and Riabov, Ivan and Rießler, Michael and Rimkutė, Erika and Rinaldi, Larissa and Rituma, Laura and Rizqiyah, Putri and Rocha, Luisa and Rögnvaldsson, Eiríkur and Roksandic, Ivan and Romanenko, Mykhailo and Rosa, Rudolf and Roșca, Valentin and Rovati, Davide and Rozonoyer, Ben and Rudina, Olga and Rueter, Jack and Rúnarsson, Kristján and Sadde, Shoval and Safari, Pegah and Sahala, Aleksi and Saleh, Shadi and Salomoni, Alessio and Samardžić, Tanja and Samson, Stephanie and Sanguinetti, Manuela and Sanıyar, Ezgi and Särg, Dage and Sartor, Marta and Sasaki, Mitsuya and Saulīte, Baiba and Savary, Agata and Sawanakunanon, Yanin and Saxena, Shefali and Scannell, Kevin and Scarlata, Salvatore and Schang, Emmanuel and Schneider, Nathan and Schuster, Sebastian and Schwartz, Lane and Seddah, Djamé and Seeker, Wolfgang and Seraji, Mojgan and Shahzadi, Syeda and Shen, Mo and Shimada, Atsuko and Shirasu, Hiroyuki and Shishkina, Yana and Shohibussirri, Muh and Shvedova, Maria and Siewert, Janine and Sigurðsson, Einar Freyr and Silva, João and Silveira, Aline and Silveira, Natalia and Silveira, Sara and Simi, Maria and Simionescu, Radu and Simkó, Katalin and Šimková, Mária and Símonarson, Haukur Barri and Simov, Kiril and Sitchinava, Dmitri and Sither, Ted and Skachedubova, Maria and Smith, Aaron and Soares-Bastos, Isabela and Solberg, Per Erik and Sonnenhauser, Barbara and Sourov, Shafi and Sprugnoli, Rachele and Stamou, Vivian and Steingrímsson, Steinþór and Stella, Antonio and Stephen, Abishek and Straka, Milan and Strickland, Emmett and Strnadová, Jana and Suhr, Alane and Sulestio, Yogi Lesmana and Sulubacak, Umut and Suzuki, Shingo and Swanson, Daniel and Szántó, Zsolt and Taguchi, Chihiro and Taji, Dima and Tamburini, Fabio and Tan, Mary Ann C. and Tanaka, Takaaki and Tanaya, Dipta and Tavoni, Mirko and Tella, Samson and Tellier, Isabelle and Testori, Marinella and Thomas, Guillaume and Tonelli, Sara and Torga, Liisi and Toska, Marsida and Trosterud, Trond and Trukhina, Anna and Tsarfaty, Reut and Türk, Utku and Tyers, Francis and Þórðarson, Sveinbjörn and Þorsteinsson, Vilhjálmur and Uematsu, Sumire and Untilov, Roman and Urešová, Zdeňka and Uria, Larraitz and Uszkoreit, Hans and Utka, Andrius and Vagnoni, Elena and Vajjala, Sowmya and Vak, Socrates and van der Goot, Rob and Vanhove, Martine and van Niekerk, Daniel and van Noord, Gertjan and Varga, Viktor and Vedenina, Uliana and Venturi, Giulia and Villemonte de la Clergerie, Eric and Vincze, Veronika and Vlasova, Natalia and Wakasa, Aya and Wallenberg, Joel C. and Wallin, Lars and Walsh, Abigail and Washington, Jonathan North and Wendt, Maximilan and Widmer, Paul and Wigderson, Shira and Wijono, Sri Hartati and Wille, Vanessa Berwanger and Williams, Seyi and Wirén, Mats and Wittern, Christian and Woldemariam, Tsegay and Wong, Tak-sum and Wróblewska, Alina and Wu, Qishen and Yako, Mary and Yamashita, Kayo and Yamazaki, Naoki and Yan, Chunxiao and Yasuoka, Koichi and Yavrumyan, Marat M. and Yenice, Arife Betül and Yıldız, Olcay Taner and Yu, Zhuoran and Yuliawati, Arlisa and Žabokrtský, Zdeněk and Zahra, Shorouq and Zeldes, Amir and Zhou, He and Zhu, Hanzhi and Zhu, Yilun and Zhuravleva, Anna and Ziane, Rayan. 2023. *Universal Dependencies 2.13*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. PID http://hdl.handle.net/11234/1-5287.

A. Appendix

Table 2 shows the most important relations going from a verb to a noun; in addition, it also shows `compound` relations going from a noun to a verb. It demonstrates that some treebanks favor the compound analysis much more than others, and three treebanks do not use the `compound` relation at all.

Table 3 shows some of the most frequent light verbs across the South Asian treebanks. Cognates are clearly observable in the Indo-Aryan languages but their preference in the individual languages varies (there are substantial differences even between Hindi and Urdu).

Language	Treebank	compound		rev. compound			nsubj		obj			iobj	obl		xcomp	conj
		NXV	**NV**	**NV**	**VN**	**VXN**	**NXV**	**NV**	**NXV**	**NV**	**VXN**	**NXV**	**NXV**	**NV**	**NV**	**VXN**
Sanskrit	Vedic						189	146	172	216	37	35	206	187	41	27
Sanskrit	UFAL						184	157	87	239	27		304	163	22	11
Hindi	HDTB	36	272		6	19	261	46	212	119	15	43	594	10	7	4
Hindi	PUD	4	90		0	1	188	18	243	269	1	33	598	9	4	3
Urdu	UDTB	36	295				189	12	230	55	9	36	472	8	8	7
Kangri	KDTB	32	199				215	72	139	115	8		334	48	16	
Bhojpuri	BHTB	152	366	2	14	33	131	21	32	51	3	11	233	9	3	5
Bengali	BRU		94				63	94	63	594	31	63	125	63		
Marathi	UFAL	5	68				341	213	130	346	3	31	252	109		8
Sinhala	STB		23	261	11	11	227	102	68	159			80	114		
Telugu	MTG		91				175	218	97	365		6	230	309		
Tamil	TTB						173	162	137	292		8	421	159		2
Tamil	MWTT			70			155	244	89	418		54	344	232	8	
Malayalam	UFAL		62	8			250	154	117	254		17	325	92	4	

Table 2: Selected relations between verbs and nouns in UD 2.13 treebanks (only main relation types are shown, subtypes are merged with their main types). The relations go from the verb to the noun except for the "reversed compound" columns, where the noun is the parent node. **NV** means that the noun immediately precedes the verb; **NXV** means that the noun precedes the verb but there are one or more words between them; analogously, **VXN** means that the verb comes first, with at least one word between it and the noun. Frequencies are shown per 10K words; an empty cell means that the relation did not occur at all while zero means that it did occur but the normalized frequency is rounded down to 0.

English	Hindi	HDTB	PUD	Urdu		Kangri		Bhojpuri		Bengali		Marathi		Sinhala		Malayalam	
do / make	karanā	118	56	krnā	170	karaṇā	48	kara	38	karā	63	karaṇe	21	kara	23	ceyyuka	25
make	karānā	3	1	krānā	1												
do / make				krūānā	1												
happen / be	honā	22	12	hūnā	37	hoṇā	76	ho/bā	72	haoŷā	31	hoṇe	10			ākuka	4
happen / be								bhaila	17								
give	denā	24	6	denā	36	deṇā	20	de	17			deṇe	3				
take	lenā	6	3	lenā	7	laiṇā	4	la	3								
apply / put	lagānā	7	1	lgānā	2			laga	3								
seem	laganā	1		lgnā	1	lagaṇā	4										
keep / put	rakhanā	2	1	rkhnā	7	rakhaṇā	4									vaykkuka	4
stay	rahanā	0		rhnā	2												
create / make	banānā	2	0	bnānā	3												
be / become	bananā	1		bnnā	1												
come	ānā	2	3	Ānā	3	āṇā	8	ā	3							varuka	4
drive	calānā	0		člānā	1												
go / walk	calanā	1	2	člnā	1			cala	8								
meet	milanā	2		mlnā	1												
show / express	jatānā	3															
pick	uṭhānā	2		āṭhānā	1	uḍāṇā	4										
cause	dilānā	1		dlānā	1												
put	ḍālanā	1		ḍālnā	1											iṭuka	8
get / find	pānā	0		pānā	2	pāṇā	4										
kill	māranā	1		mārnā	1							māraṇe	13				
fall	paḍanā	0		prnā	1	pauṇā	8										

Table 3: Selected lemmas of verbs that are connected with a noun via the `compound` relation (or its subtype), with the verb as the parent, in UD 2.13 treebanks. Frequencies are shown per 10K words; an empty cell means that the verb did not occur at all while zero means that it did occur but the normalized frequency is rounded down to 0.

Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection

Dylan Phelps[1,2], Thomas Pickard[2], Maggie Mi[2], Edward Gow-Smith[2]
Aline Villavicencio[1,3]
[1]Healthy Lifespan Institute, The University of Sheffield, United Kingdom
[2]Department of Computer Science, The University of Sheffield, United Kingdom
[3]Institute for Data Science and Artificial Intelligence, The University of Exeter, United Kingdom
{drsphelps1, tmpickard1, zmi1, egow-smith1}@sheffield.ac.uk
a.villavicencio@exeter.ac.uk

Abstract

Despite the recent ubiquity of large language models and their high zero-shot prompted performance across a wide range of tasks, it is still not known how well they perform on tasks which require processing of potentially idiomatic language. In particular, how well do such models perform in comparison to encoder-only models fine-tuned specifically for idiomaticity tasks? In this work, we attempt to answer this question by looking at the performance of a range of LLMs (both local and software-as-a-service models) on three idiomaticity datasets: SemEval 2022 Task 2a, FLUTE, and MAGPIE. Overall, we find that whilst these models do give competitive performance, they do not match the results of fine-tuned task-specific models, even at the largest scales (e.g. for GPT-4). Nevertheless, we do see consistent performance improvements across model scale. Additionally, we investigate prompting approaches to improve performance, and discuss the practicalities of using LLMs for these tasks.

Keywords: large language models, idiomaticity detection, prompting, scaling

1. Introduction

Large, pre-trained language models (LLMs) are becoming increasingly popular in academic, industrial, and lay spheres due to their ability to perform well across a range of tasks in a zero-shot or few-shot prompting set-up, including question answering, common-sense reasoning (OpenAI, 2023; Gemini Team, 2023), and machine translation (Xu et al., 2023; Koshkin et al., 2024; Dabre et al., 2023). Despite this, there is yet to be an analysis of how well such models are able to handle potentially idiomatic language. Much previous work has shown that smaller, encoder-only transformer models have poor performance in identifying and representing idiomatic expressions when pre-trained on a large general dataset (Nandakumar et al., 2019; Garcia et al., 2021). However, the performance of such models increase hugely when they are finetuned on a task-specific dataset containing a large number of idiomatic expressions (Madabushi et al., 2021; Zeng and Bhat, 2021). This fine-tuning procedure, however, requires dedicated hardware and training, something that isn't possible with LLMs on an academic budget.

In this work, we benchmark the performance of several widely-used LLMs (using both software-as-a-service remote implementations and local instances) on three in-context idiomaticity detection datasets; the idiom portion of FLUTE (Chakrabarty et al., 2022), MAGPIE (Haagsma et al., 2020), and SemEval 2022 Task 2a (Tayyar Madabushi et al., 2022). FLUTE and MAGPIE cover English (EN) only, while the SemEval dataset also includes expressions in Brazilian Portuguese (PT-BR) and Galician (GL).

Overall, our experiments show that large LLMs give competitive performance on idiomaticity datasets, which can be generally applied due to the lack of type specific fine-tuning, but nevertheless lag in general behind much-smaller finetuned encoder-only models. We also find that idiomaticity detection performance still scales with the number of parameters in the model. Finally, we discuss a number of considerations affecting the models' performance and the practicality of using them for idiomaticity detection, including the training dataset and the capability of the model to follow instructions given in the prompt.

2. Datasets

We investigate the performance of LLMs on three datasets consisting of potentially idiomatic expressions in context. The datasets are chosen to provide a diverse set of potentially idiomatic expressions which feature a range of morphological forms and variations across two different tasks: textual entailment and idiomaticity detection. 1,859 different English target expressions are represented across the three datasets. We focus on English, but the inclusion of Semeval 2022 Task 2a allows us to additionally explore performance across languages.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 178–187. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

2.1. FLUTE

FLUTE (Chakrabarty et al., 2022) frames the understanding of four kinds of figurative language (sarcasm, simile, metaphor and idioms) as a natural language inference (NLI) task, in which pairs of literal and figurative sentences are labelled as either entailing or contradicting one another. The sentence pairs are generated using a model-in-the-loop approach, with base text generated by GPT-3 which is then edited by crowdworkers and reviewed by experts.

For our analysis, we consider only the idiom section of the FLUTE dataset, which consists of 1,768 training examples across 479 idioms and a further 250 test examples across 69 idioms. No idiom appears in both the training and test sets.

Chakrabarty et al., 2022 provide benchmark performance metrics using T5 models (Raffel et al., 2020) on the FLUTE training data, reporting 79.2% accuracy (0.791 macro-F1). A FigLang22 shared task using the FLUTE dataset (Saakyan et al., 2022) attracted several entries, with the best-performing systems developed by (Gu et al., 2022) and (Bigoulaeva et al., 2022). The latter adopt a pipeline approach, improving the T5 baseline by sequentially fine-tuning on e-SNLI dataset (Camburu et al., 2018) and IMPLI (which incorporates figurative language) (Stowe et al., 2022), followed by the task dataset. Using the authors' published outputs, we calculate a macro-average F1 of 0.952 on the idiom portion of the FLUTE test set.

2.2. SemEval 2022 Task 2a

SemEval 2022 Task 2a (Tayyar Madabushi et al., 2022) is a binary classification idiomaticity detection task, in which a potentially idiomatic noun compound, as used in a given context sentence, must be labelled as either literal or idiomatic. The dataset includes compounds across a range of idiomaticity, including fully compositional (*insurance company*) as well as partially (*eager beaver*) and entirely opaque (*sugar daddy*) items. The task offers both "one-shot" and "zero-shot" settings; the former is evaluated with new context instances of previously-seen items, while the latter uses compounds not present in the training data for evaluation.

The test set for the task contains 50 compounds each in English (with 916 instances), Brazilian Portuguese (713 instances) and Galician (713 instances).

Table 1 shows the macro-F1 scores in the zero-shot and one-shot settings for the baseline models (fine-tuned multilingual mBERT, per Madabushi et al., 2021) and the best-performing entries to the shared task[1].

Setting	Reference	Language			
		EN	PT	GL	All
Zero-Shot	Best	0.902	0.828	0.928	0.890
	Baseline	0.707	0.680	0.507	0.654
One-Shot	Best	0.964	0.894	0.937	0.939
	Baseline	0.886	0.864	0.816	0.865

Table 1: Reference scores (Macro F1) for SemEval 2022 Task 2a.

2.3. MAGPIE

MAGPIE (Haagsma et al., 2020) is a corpus of instances of potentially idiomatic expressions (PIEs – expressions which have multiple senses, including at least one with a high level of idiomaticity), in which each instance has been annotated as either idiomatic, literal, or other (proper noun, etc.) by a group of crowd-sourced workers. The PIEs in the dataset are chosen from three online dictionaries and so have a wide range of forms and frequencies.

The final dataset consists of 56,622 annotated instances, of which 70% are idiomatic, 28% are literal and 1% are other. In our experiments we use the test split of the randomly split dataset, which has 4,840 instances across 1,134 PIEs).

Haagsma et al. (2020) do not provide baseline models for the MAGPIE data, but several benchmarks are provided by Zeng and Bhat (2021).

2.4. Construction Artifacts

Recent work by Boisson et al. (2023) has found that language models tuned for metaphor identification (in which they include idiomaticity detection) on artificially-constructed datasets (i.e. those not sampled from 'naturally-occurring' text) can perform well when the target expression or the surrounding context are hidden from the model, "in both cases close to the model with complete information".

As our experiments employ pre-trained LLMs without fine-tuning for the idiomaticity detection task, we anticipate that the concerns highlighted by Boisson et al. (2023) should not affect our findings. While the training regimes for many of the models we examine are not public, it seems likely that they have consumed large quantities of training data containing 'naturally distributed' idiomatic expressions.

It is also worth noting that we can not rule out the possibility that these LLMs' training data includes the training or test datasets under evaluation[2], and it is likely (for SemEval and MAGPIE) that the context sentences could have been 'seen' by the mod-

[1] For the one-shot setting, the best-performing model is a fine-tuned multilingual XLM-RoBERTa, as described in Chu et al. (2022).

[2] The SemEval test set is publicly available only without labels; FLUTE and MAGPIE are public.

els during training (albeit without idiomaticity markers), as they are taken from online sources.

3. Models

To be able to compare results from a range of currently-available LLMs, we evaluate both software-as-a-service (SaaS) and local instances of open models. To maximise applicability of our findings to researchers, we focus on local instances that can be run on consumer-level hardware (targeting a machine with 32GB RAM and 12GB VRAM).

Table 2 summarises the models used in our experiments, including the parameter count (where available), cost to run for SaaS models, and whether the training dataset is multilingual.

Model	Params (billions)	Cost ($US per 1000 tokens)	Multilingual
GPT-3.5-turbo	Unknown	0.0005	Y
GPT-4-turbo	Unknown	0.01	Y
GPT-4	Unknown	0.03	Y
Gemini-1.0 Pro	Unknown	0.000125	Y
Llama2-7B-chat	7	N/A	N
Llama2-13B-chat	13	N/A	N
Llama2-70B-chat	70	N/A	N
Phi-2	2.5	N/A	N
Mistral-7B	7	N/A	N
Flan-T5-Small	0.08	N/A	Y
Flan-T5-Base	0.25	N/A	Y
Flan-T5-Large	0.78	N/A	Y
Flan-T5-XL	3	N/A	Y
Flan-T5-XXL	11	N/A	Y

Table 2: Characteristics of the models evaluated.

3.1. Software-as-a-service Models

3.1.1. OpenAI

OpenAI models are seen to be the current state of the art in SaaS models. GPT-4 (OpenAI, 2023), their current largest model, has been shown to achieve or exceed human-level performance in a number of commonly used benchmarks. We evaluate GPT-3.5-turbo (gpt-3.5-turbo-0613), GPT-4-turbo (gpt-4-0125-preview) and GPT-4 (gpt-4) in this work. GPT-3.5 is a smaller model created as a test run during the development of GPT-4, and GPT-4-turbo is an optimised and more recent variant of GPT-4. The parameter counts for these models are not known, but it is assumed that GPT-4 is substantially larger than GPT-3.5.

3.1.2. Google

Google provides access to a number of models of varying size and price through its VertexAI API. In this work we evaluate the performance of the Gemini Pro 1.0 model. Gemini Pro is trained on a multimodal and multilingual dataset and its performance exceeds that of GPT-3.5 on a number of benchmarks (Gemini Team, 2023).

3.2. Local Models

Additionally, we evaluate the performance of popular open models that can be run locally. The models chosen are the Llama2 models, (Touvron et al., 2023) Llama2-7B-chat and Llama2-13B-chat, Phi-2 (Li et al., 2023; Abdin et al., 2023), and the CapybaraHermes[3] variant of Mistral-7B (Jiang et al., 2023).

To ensure that the models can be run on consumer-level hardware we use quantized variants of each model with 7B or more parameters. Quantization (Dettmers et al., 2022; Frantar et al., 2023) involves converting each parameter from full 16-bit floating point numbers to a set of 2^n discrete values. This massively reduces the size of the models so they can be run on a wider range of hardware, with a trade-off of lower performance. We use **Q5_K_S quantisation variants**, which use 5-bit quantization, provided by TheBloke on Huggingface[4]. 5 bit quantization has been shown to have minimal impact on the performance of the model[5].

To run the models we use the Huggingface transformers library (Wolf et al., 2020) for Phi-2 and llama.cpp[6] for all the quantized models.

3.3. Multilingual Models

We also explore the performance of multilingual models. In particular, we target our exploration to variants of the Flan-T5 models (Chung et al., 2022): Flan-T5-Small, Flan-T5-Base, Flan-T5-Large, Flan-T5-XL, and Flan-T5-XXL.

We are interested in how multilingual models' performance on idiomatic language-related tasks differs from monolingual ones. Moreover, we want to investigate the extent to which the performance is impacted by model size.

4. Results

Our main results across the three datasets (using our default prompts) are shown in Table 3. To make our results representative and generalisable, we ran the models multiple times, where not computation or cost prohibitive – all of the Flan models were run three times, whilst the Gemini Pro and GPT-3.5 models were run twice on SemEval, which is particularly important for reducing the variance of the results when testing different prompting methods; all other models were run once only.

[3] https://huggingface.co/argilla/CapybaraHermes-2.5-Mistral-7B

[4] https://huggingface.co/TheBloke

[5] See https://github.com/ggerganov/llama.cpp/pull/1684.

[6] https://github.com/ggerganov/llama.cpp

	SemEval	FLUTE	MAGPIE
GPT-3.5-Turbo	0.645	0.820	0.559
GPT-4-turbo	0.668	0.936	0.860
GPT-4	0.636	0.936	0.896
Gemini 1.0 Pro	0.672	0.924	0.721
Phi-2	0.447	0.458	0.531
Llama2 (7B-chat)	0.479	0.373	0.314
Llama2 (13B-chat)	0.505	0.602	0.483
CapybaraHermes-2.5-Mistral-7B	0.539	0.812	0.587
Flan-T5-Small	0.333	0.333	0.203
Flan-T5-Base	0.390	0.764	0.213
Flan-T5-Large	0.424	0.872	0.290
Flan-T5-XL	0.452	0.956	0.456
Flan-T5-XXL (11.3B)	0.514	0.940	0.753
baseline	0.654	0.791	0.872
best	0.890	0.952	0.955

Table 3: Main results of our models across the three idiomaticity datasets. All results presented are macro-average F1 scores over the two classes. Baseline results are taken from Madabushi et al. (2021), Chakrabarty et al. (2022) and Zeng and Bhat (2021). 'Best' results (in all cases using models fine-tuned on the task training data) are taken from Chu et al. (2022), Bigoulaeva et al. (2022) and Zeng and Bhat (2021). For SemEval, the 'zero-shot' setting is reported.

Comparing the results with the baseline and best-performing models, we can see that while the performance of large, contemporary LLMs may be higher than out-of-the-box encoder-only models, there is still a gap between them and the results which can achieved by encoders fine-tuned to the particular tasks. However, given the work of Boisson et al. (2023) on construction artifacts within datasets for idiomaticity detection, the ability of LLMs to disambiguate a wide-range of PIEs without additional fine-tuning shows the general ability of these models to detect idiomaticity, which may not have been achieved by fine-tuned encoders.

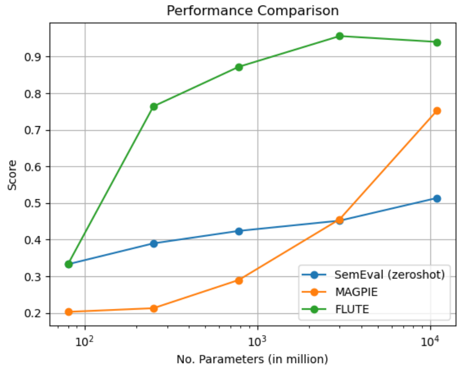


Figure 1: Performance on the three datasets for different Flan-T5 model sizes.

4.1. Model Scaling

With the exception of the Mistral-7B model, there is a significant gap in performance between the smaller, locally-run models and the larger SaaS models. We can also see the same trend for our Llama2 models, where the larger Llama2-13B model outperforms the smaller Llama2-7B one on all datasets and splits. From the results of the Flan-T5 model variants, as shown in Figure 1, there is a clear trend that increasing model size leads to improved performance. This trend appears to slow down somewhat after model size reaches around 3B parameters (Flan-T5-XL), though performance on the MAGPIE dataset continues to grow.

4.2. Prompts

Due to the differing input formats required by the various models, we use slightly different prompts. Here, we show our default prompts used for the GPT models. For SemEval and MAGPIE, we use:

"Disambiguate whether the given expression is used idiomatically or literally in the given context, returning 'i' if the expression is being used idiomatically or 'l' if literally. Expression: <PIE>. Context: <target sentence>. Only return one letter (i or l)."

For the FLUTE entailment task, we use:

"Disambiguate whether the second sentence follows from the first, returning 'entailment' if it does, and 'contradiction' if not. Sentence 1: <premise sentence> Sentence 2: <hypothesis sentence>."

	EN
Default	0.739
“Expert in language use”	0.635
“Expert in language use” + Idiomatic vs. Compositional	0.717
“Expert in Idiomatic Language”	0.538
No “Only return one letter (i or l).”	0.633

Table 4: Results (macro F1) on the English test set of SemEval with GPT-3.5-turbo using prompt engineering.

	GPT-3.5-turbo		Gemini 1.0		Flan-T5-XXL	
	PT	GL	PT	GL	PT	GL
Default	0.553	0.587	0.582	0.604	0.464	0.411
Language Prompt	0.554	0.604	0.561	0.640	0.479	0.457
Translated	0.541	0.512	0.549	0.665	0.573	0.477

Table 5: GPT 3.5-turbo, Gemini 1.0, and Flan-T5-XXL results for Portuguese and Galician on SemEval using multilingual prompts.

4.3. Prompt Engineering

We investigate the effect of several prompt variations on performance for GPT-3.5-turbo on the English SemEval test set. As part of the OpenAI API, there are two prompts: “system” and “user”. We first tried using the system prompt to define the task for the model, but obtained better performance using only the user prompt – this aligns with the experiences of others that GPT-3.5 often doesn’t follow the system prompt well, unlike GPT-4[7].

We present our results for this in Table 4. Note that variation between runs using the same prompting strategy is high (up to 0.04 F1), which leads to difficulty in discerning the effect of changing the prompt.

Expert impersonation is motivated by work which has shown that prompting LLMs to impersonate domain experts can lead to higher performance (Salewski et al., 2024). As such, we tried two approaches; starting the prompt with “You are an expert in language use.” or “You are an expert in idiomatic language.”. However, we find that neither of these approaches lead to improved performance. Interestingly, replacing the word “Literal" with “Compositional" did seem to have a positive effect. We found that removing the instruction to explicitly return only one letter (‘i’ or ‘l’) led the model to occasionally return other outputs, which causes a drop in performance (as we treat such responses as invalid). For the English subset, this is the case for 3% of outputs (28 out of 916 examples).

4.3.1. Language Prompts

Since SemEval has test data in English, Portuguese, and Galician, we experiment with a) explicitly stating the language of the sentence in the prompt, and b) translating the prompt using a commercial machine translation tool. We perform this analysis for GPT-3.5-turbo, Gemini 1.0 Pro, and Flan-T5-XXL, with results shown in Table 5.

For Gemini 1.0 Pro and Flan-T5-XXL we see performance improvement for Galician under both of these approaches, with higher performance when translating the prompt. We hypothesise that both English and Portuguese are likely well-represented in the model training data, and LLMs in general work well in multilingual settings (Shi et al., 2022). However, Galician is likely to be both rare and potentially confused with Portuguese when the language is not specified, or when there is less text in that language available in the prompt. It would be interesting to experiment further with similar language pairs.

Not shown here is that we recorded reduced performance for English across all three models when specifying the language in the prompt (0.739 to 0.674 for GPT-3.5-turbo, 0.771 to 0.732 for Gemini 1.0 Pro, 0.716 to 0.706 for Flan-T5-XXL). It is possible that additional prompt tokens specifying the language may act as a ‘distractor’ when it is the *de facto* default, and the nature of the generative models means that we can anticipate variation in responses to identical prompts.

4.4. Few-shot Prompting

The “one-shot” setting of SemEval 2022 Task 2a (in which further examples of the target PIE in context are made available) allows for the investigation of passing examples to the model through the prompt. We thus experiment with doing so for GPT-3.5-turbo, Gemini 1.0 and Flan-T5-XXL. We try two configurations: passing one example per PIE (one-shot), and passing all the examples that are available in the dataset (few-shot)[8]. These results are shown in Table 6.

Interestingly, the impact of few-shot prompting varies across the models. Flan-T5-XXL benefits the most from this, with stark and consistent performance improvements across the three settings and across all three languages – the overall F1 jumps from 0.580 in the Zero Shot setting to 0.805 in the Few Shot setting.

Further to this we analyse the performance of all size Flan-T5 models, and present a heatmap illustrating the impacts on performance stemming from zero-shot and few-shot scenarios in Table 7.

[7] https://community.openai.com/t/what-is-the-difference-between-putting-the-ai-personality-in-system-content-and-in-user-content/194938

[8] Where available, the one-shot training data has one idiomatic example for each PIE, and one literal example. However, for some PIEs just one of these is present.

Model	Setting	EN	PT	GL	All
Gemini Pro 1.0	Zero-shot	**0.766**	0.590	0.600	0.672
	One-shot	0.706	0.625	0.711	0.688
	Few-shot	0.685	**0.642**	**0.745**	**0.693**
GPT-3.5-turbo	Zero-shot	**0.739**	**0.563**	**0.579**	**0.645**
	One-shot	0.645	0.542	0.553	0.594
	Few-shot	0.686	0.545	0.566	0.614
Flan-T5-XXL	Zeroshot	0.629	0.464	0.411	0.514
	Oneshot	0.810	0.665	0.732	0.749
	Fewshot	**0.845**	**0.713**	**0.828**	**0.805**
Best	Zero-shot	0.964	0.894	0.937	0.939

Table 6: Results on SemEval using few-shot prompting.

	Small	Base	Large	XL	XXL
Oneshot (EN)	0.432	0.079	0.199	0.348	0.182
Oneshot (PT)	0.388	0.011	0.227	0.228	0.202
Oneshot (GL)	0.526	0.049	0.053	0.185	0.321
Oneshot (ALL)	0.443	0.054	0.162	0.264	0.235
Fewshot (EN)	0.516	-0.003	0.332	0.404	0.217
Fewshot (PT)	0.391	0.000	0.093	0.285	0.249
Fewshot (GL)	0.576	0.000	0.137	0.354	0.417
Fewshot (ALL)	0.489	-0.001	0.227	0.352	0.291

Table 7: Enhancements in Macro F1 scores (positive values) and declines (negative values) when compared to the performance in zero-shot conditions across all Flan-T5 models.

The smallest models benefited the most from seeing one or more examples before inference. In the best cases, performance in English improved by 0.432 in the one-shot setting and 0.516 in the few-shot setting. Interestingly, few-shot prompting can be seen to improve performance across Portuguese and Galician examples in all model settings, apart from T5-FLAN-Base and Large where there is little, or no improvement. It appears that Flan-T5-Base seems to be least improved by prompting with examples, with a negative effect on performance in few-shot prompting settings. In the one-shot setting, improvement in model performance is minor. The Large, XL and XXL models also benefited from one- and few-shot prompting, with Flan-T5-XL seeing the most performance enhancement. It appears that whilst models follow “bigger is better" in zero-shot settings, they do not necessarily follow this pattern under one/few-shot prompting. In fact, the best performance in the few-shot setting is with T5-Small, which at only 80M parameters achieves an overall F1 of 0.821, the best performance of any of the models we have evaluated in this paper. This is in significant contrast to performance on MAGPIE and FLUTE, where zero-shot performance is very low. The model is likely learning some artefacts from the data such as predicting only one label for a given PIE in the SemEval dataset.

Gemini 1.0 Pro also achieves consistent (though smaller) performance improvements from Zero Shot to One Shot to Few Shot, but the performance for English reverses this pattern. We also see a big jump in performance between Zero Shot and One Shot for Galician, which we again attribute to the rarity of this language and its similarity with Portuguese.

GPT-3.5-turbo is hindered by providing examples. The reasons for this are unclear, but this may be linked to the inability shown by GPT-3.5 to follow system prompts. If the model is not successfully following longer prompts then they may effectively introduce noise and lead to worse performance, as we saw when comparing results with and without system prompts.

5. Discussion

5.1. Task Labelling

The majority of the models we examined achieved high performance on the FLUTE dataset. We attribute this to the nature of FLUTE's evaluation being distinct from MAGPIE and SemEval. For the latter two, the model is asked to label ‘idiomatic’ or ‘literal’ use of a given idiom, whereas, in the FLUTE STS task, the model is required to pick out the contradiction or entailment relationship between two sentences.

This means that a model might not necessarily require ‘knowledge’ of the target idiom to succeed, but could determine the relationship between the two sentences from other information, as facilitated by contextualised embeddings (Boisson et al., 2023). Moreover, the model is likely to have encountered similar tasks during its pre-training. Flan-T5 models are instruction-refined versions of T5 (Raffel et al., 2020; Chung et al., 2022), that have undergone exposure to over 1000 tasks during its fine-tuning process alone. Among these tasks are evaluations of entailment and contradiction judgments, akin to FLUTE, such as SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), CB (de Marneffe et al., 2019) and numerous other reasoning tasks (for details see Raffel et al., 2020; Chung et al., 2022).

5.2. Practicalities

In contrast with fine-tuned classification models, as prompted models are capable of open-ended generation, they may not output a response in the format requested. While the output may be readily interpretable by a human reader, this is not practical when evaluating large numbers of responses. Prompting for specific formats is easier for models which have undergone more instruction tuning (Ouyang et al., 2022; Rafailov et al., 2023), and is a key reason why the Mistral-7B model outperforms the Llama2 7B variant.

Prompted, generative models produce outputs which are subject to variation when they are repeatedly given the same prompt. While the user may have some control over this behaviour through 'temperature' parameters, this variability is inherent to generative models. When converting the outputs of such models to a labelling decision, this variability will also affect the results.

Despite their generally higher performance than the local models and their advantages when it comes to prototyping, there are a number of considerations specific to SaaS models which may be significant. These include:

1. Cost – The larger models have a higher per-1000-tokens cost, which may lead to some evaluations being cost-prohibitive. Evaluating GPT-4 on the (relatively small) SemEval test set, for example, costs $11. Running evaluation on this model, especially across multiple runs for prompt tuning, etc. may potentially price out researchers with lower budgets.

2. Safety Features – Commercial SaaS models frequently include features designed to limit models and users' capability to process or generate content which may cause harm. These features may also impact on researchers' ability to use the tools, as they produce what are effectively false positives. For example, when using the VertexAI API for experiments with Gemini Pro, the API consistently refused to generate responses for a small number of prompts. These included certain contexts for the expression *street girl* which referred to prostitution or sexualization, but also the FLUTE sentence pair "Your brother is mature and behaves in an adult manner. Your brother is a big baby." for the expression *to be a big baby*[9]. We treat any such responses as incorrect in our statistics.

3. Service Changes – Changes to the underlying model can be made by the third party at any time, and can significantly impact the performance of the models and the consistency of results. Whilst undertaking this work the default gpt-3.5-turbo model changed from one released in June 2023, to one released in January 2024.

4. Rate limits – For larger datasets, the rate limits of commercial APIs can become an issue. As it is still not fully released, for a significant amount of time during the creation of this work, the daily rate limit for GPT-4-turbo was lower than the number of tokens in MAGPIE, which prevented us from completing any evaluation runs for this model and dataset combination.

6. Conclusion

In this work we have evaluated the performance of various large language models on three idiomaticity datasets (SemEval 2022 Task 2a, FLUTE, and MAGPIE). We have investigated locally-run models up to 13B parameters, as well as significantly larger models (GPT-3.5, GPT-4, and Gemini 1.0 Pro) accessed through commercial APIs. We perform an extensive analysis of the impact of several factors on performance; model size, prompt engineering and few-shot prompting. In addition, we discuss considerations for practitioners wishing to use these models in their own work, with emphasis on cost and practicalities such as the variability of outputs and the impacts of decisions made by the companies operating these services. Our overall findings are as follows: 1) LLMs at the highest scale are able to achieve competitive results for idiomaticity detection, and performance on FLUTE in particular seems to have saturated, but these general models do not match the performance of (much-smaller) encoder models fine-tuned for the specific idiomaticity detection tasks of SemEval and MAGPIE. 2) The performance of prompted, generative LLMs seems to scale consistently with parameter count for these datasets, indicating the potential of even bigger models to achieve further increases in performance. 3) While they are based on a relatively small set of examples, our experiments with multilingual models suggest that performance gains can be obtained by specifying the target language, translating prompts and by providing examples. However, the efficacy of these modifications depends on the model used and the language in question; they appear to harm performance for English (which is, presumably, the most-represented language in the model training regimens) while producing the largest benefit for the much rarer Galician.

Acknowledgments

We would like to thank the following sponsors for supporting this work:

- The Healthy Lifespan Institute (HELSI) at The University of Sheffield, UK EPSRC grant EP/T517835/1,

- The Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1], and

- UK EPSRC grant EP/T02450X/1, by the Turing Institute Fellowship and by the UniDive COST Action.

[9] Replacing the word 'adult' with 'grown-up' convinced the service to generate a response.

7. Bibliographical References

Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Mojan Javaheripi, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. 2023. Phi-2: The surprising power of small language models.

Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, Aline Villavicencio, and Iryna Gurevych. 2022. Effective Cross-Task Transfer Learning for Explainable Natural Language Inference with T5. ArXiv:2210.17301 [cs].

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction Artifacts in Metaphor Identification Datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590, Singapore. Association for Computational Linguistics.

Zheng Chu, Ziqing Yang, Yiming Cui, Zhigang Chen, and Ming Liu. 2022. HIT at SemEval-2022 Task 2: Pre-trained Language Model for Idioms Detection. ArXiv:2204.06145 [cs].

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Raj Dabre, Bianka Buschbeck, Miriam Exel, and Hideki Tanaka. 2023. A Study on the Effectiveness of Large Language Models for Translation with Markup. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 148–159, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics*, pages 3551–3564. Association for Computational Linguistics (ACL).

Google Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Gregori Gerganov. 2024. Llama.cpp: Inference of meta's llama model (and others) in pure c/c++.

Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. 2022. Just-DREAM-about-it: Figurative Language Understanding with DREAM-FLUTE. ArXiv:2210.16407 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024. TransLLaMa: LLM-based Simultaneous Translation System.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks Are All You Need II: phi-1.5 technical report.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. *arXiv preprint arXiv:2109.04413*.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34.

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex

Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A Report on the FigLang 2022 Shared Task on Understanding Figurative Language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2024. In-context impersonation reveals large language models' strengths and biases. *Advances in Neural Information Processing Systems*, 36.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic Expression Identification using Semantic Compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey.

8. Language Resource References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative Language Understanding through Textual Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287. European Language Resources Association.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI Models' Performance on Figurative Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Universal Dependencies for Saraiki

Meesum Alam, Francis M. Tyers, Emily Hanink, Sandra Kübler
Indiana University
{meealam,ftyers,emhanink,skuebler}@iu.edu

Abstract

We present the first treebank of the Saraiki/Siraiki [ISO 639-3 skr] language, using the Universal Dependency annotation scheme (de Marneffe et al., 2021). The treebank currently comprises 587 annotated sentences and 7 597 tokens. We explain the most relevant syntactic and morphological features of Saraiki, along with the decision we have made for a range of language specific constructions, namely compounds, verbal structures including light verb and serial verb constructions, along with different types of relative clauses.

Keywords: Saraiki, Universal Dependencies, Indo-Aryan Languages

1. Introduction

Universal Dependencies (UD) is now a widely used annotation scheme for developing syntactic annotations and parsers for a language (de Marneffe et al., 2021; Nivre and Zeman, 2020). It already covers around 220 languages around the world and is growing rapidly. These linguistically annotated corpora are crucial sources for NLP projects of any language. However, Indo-Aryan languages have received little attention in both UD and NLP applications. There currently exist Universal Dependency treebanks for Hindi (Ravishankar, 2017), Urdu (Ehsan and Butt, 2020), and Punjabi (in Gurmukhi script) (Arora, 2022). No lesser studied Indo-Aryan languages are covered in the UD project.

We present a UD treebank for Saraiki, a language of 25 million speakers, which is considered a neglected language in Pakistan. We follow the existing UD guidelines for the annotation where possible. Here, we describe our decisions for phenomena specific to the Saraiki language.

The remaining sections are as follows: Section 2 provides background on the Saraiki language, Section 3 discusses work on treebank construction for related languages, and Section 4 describes the corpus and annotation process. Section 5 discusses part of speech and morphological characteristics of those word classes necessary to understand the discussion of language specific phenomena, and Section 6 discusses the decisions made for language specific phenomena, namely compounds, verbal structures including light verb and serial verb constructions, as well as different types of relative clauses.

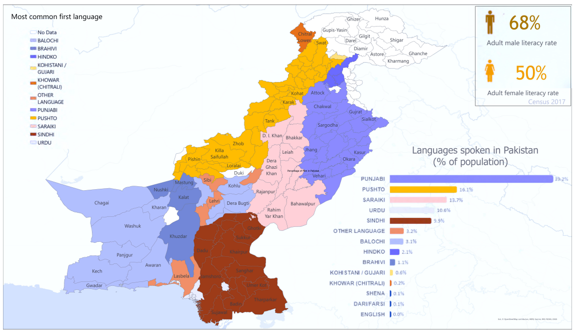


Figure 1: Map showing the percentage and distribution of languages in Pakistan. The region where Saraiki is spoken is shown in pink.

2. Saraiki

Saraiki is an Indo-Aryan language widely used in Pakistan and India. The language is one of the ancient languages of the region. Saraiki is spoken by around 25 million people in Southern and Southwestern Punjab and Northern Sindh (see the map in Figure 1). Saraiki is also known as Jataki, Multani, Thali, Riasti and Deraywal in various regions of the Punjab. Saraiki, also spelled *Siraiki*, is counted among the widely-spoken languages in the Pakistani provinces of Punjab and Khyber Pakhtunkhwa (KPK). It is the sister language of Punjabi and Sindhi but has not received much attention in linguistics research.

Saraiki is written from right to left in Perso-Arabic script. It is head-final and follows a basic Subject-Object-Verb (SOV) structure within clauses. According to Bashir and Conners (2019), Saraiki word order is relatively free: Topic and focus marking are generally achieved by changes in word order. Saraiki does not have definite or indefinite markers, but it does have numeric ہِک (*hik* ‘one’) to mark indefiniteness. Saraiki is a pro-drop language, it uses clitics/pronomial suffixes in perfective transitive sentences to mark the subjects on verbs. Saraiki has split ergative alignment in addition nominative-absolutive alignment. For more details, see section 6.2.1.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 188–197. May 25, 2024.

Source	Sentences		Tokens	
	Untagged	Tagged	Untagged	Tagged
Common Voice (Ardila et al., 2020)	5 712	288	52 300	17 500
Jhok Newspaper (Dhareja, 2017–2022)	56 000	177	1.15M	5 700
Linguistic examples	—	122	1 851	1 851

Table 1: Textual basis of the Saraiki Treebank.

Saraiki shares morphological and syntactic features with Punjabi but differs on the phonological level, which has allowed it to evolve into a distinct but related language (Bashir and Conners, 2019). As the language has been spoken in different regions of Pakistan for a long time, multiple dialects have emerged over time. Shackle (1976) distinguishes six varieties: Southern Sararik, Northern Saraiki, Sindhi Saraiki, Jhangi Saraiki.

3. Related Work

NLP applications heavily rely on linguistically annotated resources; these resources have multiple functions as they test the linguistic theories, are used to train and evaluate parsing technologies, and provide insights into specific linguistic phenomena of a language (Nivre and Zeman, 2020). However, the Indo-Aryan (IA) languages lack good digital tools because of the scarcity of available corpora. This is also true for Universal Dependency treebanks; we find some IA languages added to the repository. These treebanks cover the major languages: Hindi (Tandon et al., 2016), Urdu (Bhat and Sharma, 2012), Marathi (Ravishankar, 2017), and Punjabi (Arora, 2022). Additionally, there are automated conversions of Urdu (Ehsan and Butt, 2020) and Hindi (Bhat et al., 2018) treebanks from constituent annotations.

For Saraiki, there is little research in the area of NLP. Alam et al. (2023) have developed a morphological analyzer for Saraiki, and Asghar et al. (2021) created a part of speech (POS) tagger. There is also ongoing work on a Saraiki wordnet under Higher Education of Pakistan's Funding at Sarghoda University (Gul et al., 2021), but the system has not been released yet. For the development of NLP related tools, it is equally important to understand the linguistics phenomenon of a language; Bashir and Conners (2019) have published a descriptive grammar for Saraiki, which we used as the basis for our treebank annotations.

4. Corpus and Annotation Process

The Saraiki treebank currently consists of 587 sentences, corresponding to 7 597 tokens in total.

Our treebank is based on sentences from three different sources: from the Saraiki Common Voice corpus (Ardila et al., 2020), from the Jhok newspaper (Dhareja, 2017–2022)[1], and sentences generated during the annotations discussions, to clarify decisions on specific syntactic phenomena in Saraiki. Table 1 shows the distribution of the different text types. Saraiki is under-resourced language and it is difficult to find digital texts in this language, thus limiting our options in creating a diverse textual basis for the treebank.

In a first step, the data was converted into CoNLL-U format and manually segmented. The data have been shared with Saraiki speakers and linguistics scholars in Pakistan. This helped in making decisions on parts of speech (POS) tagging. We manually annotated the corpus for parts of speech. Since there does not exist a standard POS tagging scheme for Saraiki, we left the XPOS category for future work. The POS tagged text was used for the development of a Saraiki morphological analyzer (Alam et al., 2023). Then we started annotating the corpus for universal dependencies. We currently have 587 sentences fully annotated, and will add more annotations in the future. Once we reach 1 000 sentences, the treebank will be published via the UD project.

The annotation is carried out in two steps by the first author, a native speaker of Saraiki, in consultation with the other authors. For part of speech tagging, difficult cases are resolved based on information from the the Saraiki dictionary (Jukes, 2019), along with consulting Saraiki speakers and experts from the Urdu Universal Dependency Treebank to validate decisions. The dependency relationships are annotated using Annotatrix (Tyers et al., 2017), in consultation with all co-authors and UD experts.

5. Saraiki Parts of Speech and Morphology

As of today, there does not exist a language specific part of speech tagging scheme for Saraiki. Even though there are schemes for Punjabi (Gill et al., 2009) and Urdu (Hardie, 2003), we forcused on the Universal POS tagset (Petrov et al., 2012), leaving the XPOS category for future work. All of the UD POS tags occur in our corpus; Table 2

[1] These sentences are used with permission from the newspaper.

POS Tag	Count	Percent
NOUN	1314	17.3
VERB	1231	16.2
PUNCT	759	10.1
ADJ	714	9.4
ADP	630	8.3
PRON	569	7.5
ADV	501	6.6
PROPN	417	5.5
AUX	387	5.1
CCONJ	386	5.1
DET	258	3.4
SCONJ	190	2.5
PART	188	2.5
INTJ	22	0.3

Table 2: Distribution of Universal Dependency parts of speech tags in the Saraiki Treebank.

gives a detailed picture of the distribution of the tags in the Saraiki Treebank.

Verbs Similar to other Indo-Aryan languages, Saraiki verbs undergo derivational and inflectional processes. Saraiki verbs inflect for number, gender, tense, aspect, and mood. Adverbs, compounds, and reflexives can be derived from verbs via derivational verbal morphology. Additionally, Saraiki uses verb stem alteration. To describe those, we use work by Bashir and Conners (2019) on the eight different verb stem alterations as the basis for our annotations.

In Saraiki, certain verbs play a dual role. When occurring within a light verb construction, they take the role of auxiliaries, providing information on the verb's aspect. Consequently, we distinguish between VERB and AUX, according to the structure. For infinitives, we follow decisions in the Punjabi treebank (Arora, 2022): We mark them as VERB in all instances, regardless of their semantic interpretation.

Nouns We found three types of nouns in our treebank: case-marked nouns, non case-marked nouns, and uninflected nouns. Most nouns are case-marked in addition to being inflected for gender and number. Saraiki uses four cases: direct, oblique, vocative, and ablative. Examples of nouns that can be case-marked are ماں (*maa'n* 'mother') and چھاں (*chaa'n* 'shade'). The second type of nouns are non case-marked nouns. These nouns are borrowed from neighboring languages, and are adapted to suit Saraiki morphology. Examples of this type are ٻال (*baal* 'male child') and ذات (*zaat* 'caste'). The last category of nouns does not take any kind of inflections; these nouns are mostly borrowed from Urdu or Persian, such as ایمان (*emaan* 'faith') and رب (*Rub* 'God').

Adjectives In Saraiki, adjectives take the case and inflection of the nouns that they modify. If a noun is not case-marked, modifying adjectives agree with it in gender and number only.

Pronouns and demonstratives Saraiki does not distinguish between third person proximal and distal pronouns and demonstratives. Instead, the distal forms for *he, she, that, those* اوں (*oo'n*) are used for both expressions alongside their proximal forms اے (*ay* 'he, she, it, this, these').

Following Bashir and Conners (2019), who identify a morphological difference between relative pronouns that stand alone or immediately precede a noun, we annotated relative pronouns as PRON where they function as independent pronouns and DET where they function as determining adjectives. The adjectival forms, unlike the stand-alone pronominal forms, inflect robustly for number, gender, and case of the noun they precede and modify.

6. Annotation Decisions

In this section, we focus on language specific constructions, focusing on the treatment of (split) ergative sentences, serial and light verbs, as well as compounds and relative clauses. Remember that Saraiki is head-final and written right to left.

6.1. Compounds

Saraiki has a comprehensive system of creating multiword expressions and compounds in open and closed POS categories. In section 6.2, we will focus on the V-V compound in serial verb and light verb constructions. Here, we discuss an additional type of V-V compounding, reduplication, plus compounds involving nouns, reflexive pronouns, and adverbs.

Reduplication This is common for emphasis, for noun compounding and pluralization. In these cases, we annotate the verbs using `compound:redup`, with the first verb as the head. Interestingly, reduplication can occur with all open class categories. Verb reduplication is different from light or serial verb constructions. These verbs do not provide tense, aspect, and modality information, and they are not part of complex serial verb predicates. In example (1), گھت (*ghut* 'put') is reduplicated, either for emphasis or to indicate a quick action. As described above, reduplication can be used with almost all open categories of the

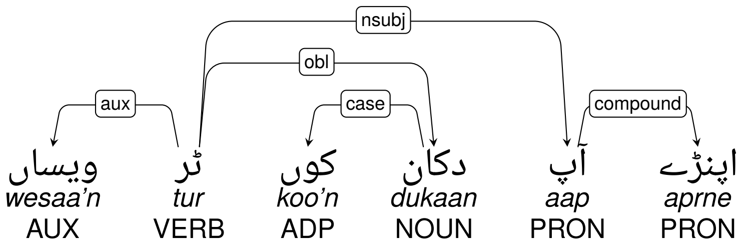
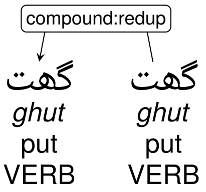


Figure 2: The annotation of the example in (4).

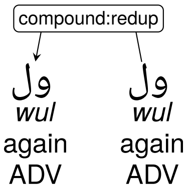
grammar in Saraiki. In example (2), reduplication is used to emphasize the adverb ول (*wul* 'again').

(1)



"put quickly"

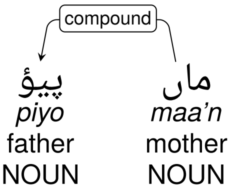
(2)



"**Again**" (emphasized)

Noun-Noun Compounds In Saraiki, there are a wide range of concepts that are expressed as noun-noun compounds. We use the `compound` relation in these cases. Example 3 shows a combination of ماں (*maa'n* 'mother') and پیؤ (*piyo* 'father') meaning "parents".

(3)



"parents"

Reflexive Pronouns These are constructed by combining the two words اپنڑے (*apnre* 'own') and آپ (*aap* 'self') in a multi-word expression (see example 4 and Figure 2). We follow the UD guidelines and use the `compound` relation to combine those two words.

(4) ویساں ٹر کوں دکان آپ اپنڑے
wesaa'n tur koo'n dukaan aap aprne
AUX.FUT go to shop-ACC PRON PRON
"I will go to the shop by myself"

6.2. Verbs

In Saraiki, the verb system is more complex than in the neighbouring languages Punjabi, Urdu, and Hindko (Bashir and Conners, 2019). Syntactically, Saraiki exhibits split ergativity in addition to pronominal suffixation onto verbs in some contexts. It uses two types of light verb constructions: one consisting of two verbs where one verb acts as an auxiliary, contributing only tense, aspect and modality information, and another consisting of a noun or adjective in addition to the light verb. Additionally, Saraiki employs serial verb constructions. We will discuss all these phenomena and annotation decisions in more detail below. In the Common Voice corpus by Ardila et al. (2020), out of all the verbs construction we found approximately 21% light verb constructions; interestingly, half of these light verb constructions use the verb تھیوݨ (*thivaṇ* 'to become'). These numbers are based on the current treebank, but we expect the percentages to remain stable as we add more sentences.

6.2.1. Syntactic Split Ergativity

Saraiki belongs to the group of languages that have both nominative–accusative and ergative-absolutive alignment (see Dixon (1994) for an overview). According to Bashir and Conners (2019), Saraiki shows an ergative-absolutive pattern only in perfective contexts, a pattern common across Indo-Aryan languages. It is important to know that unlike Urdu, Punjabi, and Hindi, Saraiki lacks a dedicated ergative morpheme. Consequently, the effects of this split are observable only in verbal agreement patterns. The generalization is that verbs agree with agents of transitive verbs and subjects of intransitive verbs in the same way in the imperfective aspect, but do not agree with agents of transitive verbs in the perfective aspect. Thus, while patients are oblique in imperfective contexts, it is agents that are oblique in perfective contexts. Table 3 lays out the case alignment pattern across imperfective and perfective contexts.

The aspectual contrast giving rise to this split is exemplified below. The imperfective sentence in example (5) shows a typical nominative-accusative agreement pattern, in which the verb

	Intransitive	Transitive	
	Subject	Agent	Patient
Perfective	Nom	Obl	Nom
Imperfective	Nom	Nom	Obl

Table 3: Split-ergative alignment in Saraiki. Subjects of intransitive verbs are always nominative, while agents and objects of transitive verbs depend on the aspect of the verb. In perfective aspect, the oblique encodes the agent, while in imperfective aspect the oblique encodes the patient.

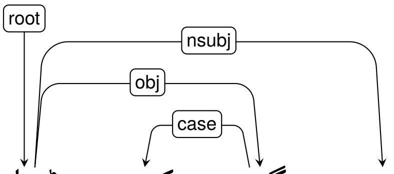
agrees with the nominative argument قاسم (*Qasim* 'Qasim'). The same case and agreement pattern is found with intransitive verbs, which agree with their nominative subject.

In the perfective sentence in example (6) in contrast, the agent of the transitive verb پڑھی (*parhi* 'read'), قاسم (*Qasim* 'Qasim') carries the oblique case, while the direct object کتاب (*kitaab* 'book') carries nominative case. Notably, the verb in this context agrees with its direct object rather than its subject. The generalization is thus that, in perfective contexts only, agents of transitive verbs i) are oblique arguments ii) may not control subject agreement.

(5) پڑھدا اے کتاب
ay parhda *kitaab*
AUX read-PRES-SG-M book-OBL-SG-F
قاسم
qasim
Qasim-NOM-SG-M
"Qasim reads the/a book"

(6) پڑھی ہا کتاب
ha parhi *kitaab*
AUX read-PP-SG-F book-SG-F
قاسم
qasim
Qasim-OBL-SG-M
"Qasim read the/a book"

In our treebank, both patterns are present. For the ergative sentences, we decided to follow the Urdu (Ehsan and Butt, 2020) and the Hindi treebank (Bhat and Sharma, 2012), we annotate agents as `nsubj` and patients and other non-agents as `obj`. We are aware that this does not agree with the decisions made in the Basque treebank (Aduriz et al., 2003), which uses `subj` for such arguments in the ergative.

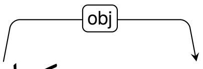


(7)
ڈھایا کوں گھر اوں
dahaya *koo'n* *ghar* *oo'n*
demolished to house-ACC he
VERB ADP NOUN PRON
"He demolished the house"

Example (7) shows an example of an ergative sentence, where we annotate the agent اوں (*oo'n* 'he'), which is in the oblique case, is the subject, and گھر (*ghar* 'house') is the direct object in ergative case.

We note that another type of agent marking is also available. This strategy uses pronominal suffixes (clitics) on the verb to mark the grammatical features of the agent. In this type of structure, the transitive verb in the perfective form shows object agreement, with the pronominal agent cliticized onto the end of the verb. In example (8), the verb پیتم (*pita-m* 'I drank') agrees with the noun پاݨی (*paanri* 'water'), and the agent 1.M.SG is added to the end of the verb پیتم. In example (9), the verb کھادئیس (*khād-i-s* 'he ate') agrees with بھاجی (*bhaj-i* 'food-F.SG'), and the agent is marked on verb.

(8) پیتم پاݨی
pita-m *paanri*
drink-PST-1.M.SG water.M.SG
"I drank water"



(9)
کھادئیس بھاجی
khā-d-i-s *bhaj-i*
eat-PERF-F.SG-M.3SG food.F.SG
VERB NOUN
"He ate food"

These constructions are possible only in the perfective forms. Note that while Bashir and Conners (2019) call these pronominal suffixes, Syed and Raza (2019) call them clitics. On either treatment, this type of construction is sensitive to the morphological features of the agent, which are marked on the verb. Following the UD guidelines, we annotate the argument as direct object `obj`.

This morphologically embedded ergativity (differential case marking) is also found in Hebrew (Glinert, 2004) and Hungarian (Bárány, 2012).

6.2.2. Serial Verb Construction

Serial verbs mostly conceptualize one event and are realized as one linear, complex predicate with-

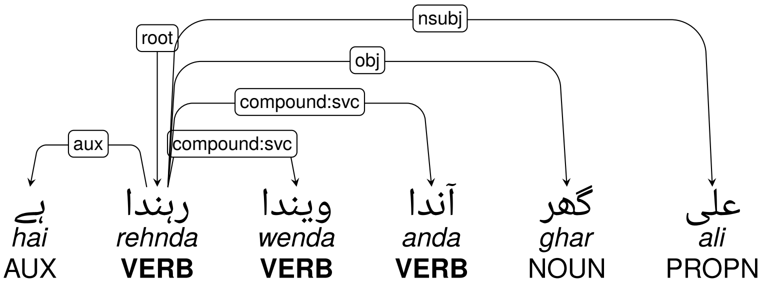


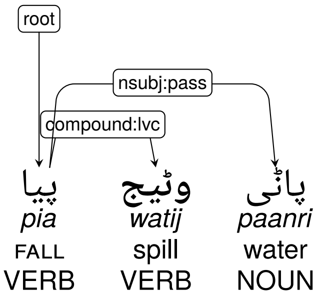
Figure 3: The annotation for the serial verb construction (POS of serial verbs in bold) of example (10).

out explicit coordination or subordination markers. This feature is common in many IA languages. Example (10) shows a sentence from our treebank, and Figure 3 shows our annotation. Since we do not yet know enough about the constraints on this construction, we decided to annotate the involved verbs serially. As Saraiki is a head final language (written from right to left), we mark the last verb as the head of the clause and create `compound:lvc` relations with other verbs. We anticipate changes to these annotations in the future once we have a better understanding of this construction.

(10) علی گھر آندا ویندا رہندا ہے
hai rehnda wenda anda ghar ali
AUX keeps go come home Ali-NOM
'Ali keeps coming and going home"

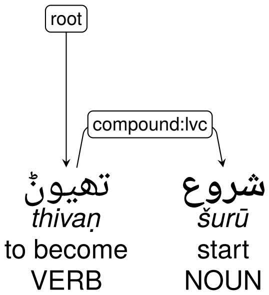
6.2.3. Light Verb Constructions

In Saraiki, we find sequences of verbs where the main verb is followed by another 'light' verb, in addition to constructions in which a light verb is followed by a noun or adjective. In both cases, the light verb has little semantic content. In V-V LVCs, the second verb mostly contributes information about aspect or modality. All such constructions have been given the dependency of `compound:lvc`. We show an example in (11): وٹیج (*watij* 'spill') is the main verb in the structure, and پیا (*pia* 'fall') provides aspectual information about the main verb, indicating that the action is completed.

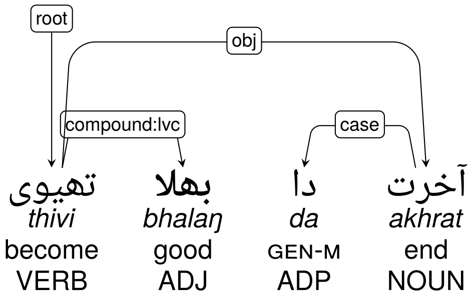


(11)
"The water was spilled"

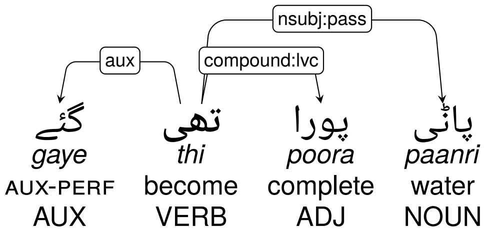
In the treebank, we also found the verb تھیوݨ (*thivaṇ* 'become'), a change of state verb (Bashir and Conners, 2019) in Saraiki, which, unlike ہووݨ (*hovaṇ* 'be'), appears in SVCs, LVCs, and as an auxiliary. تھیوݨ (*thivaṇ* 'become') can also be followed by another light verb construction. Where it occurs in a light verb construction, we mark it as a root with a `compound:lvc` dependency to the noun or verb (see examples (12) and (13)); when تھیوݨ (*thivaṇ* 'become') is not part of the light verb construction, we mark it as an auxiliary AUX (see example (14)).



(12)
"to start"



(13)
"may (you) have a better end"



(14)
"The (land) filled (with) water" (lit.: water full become go-PERF)

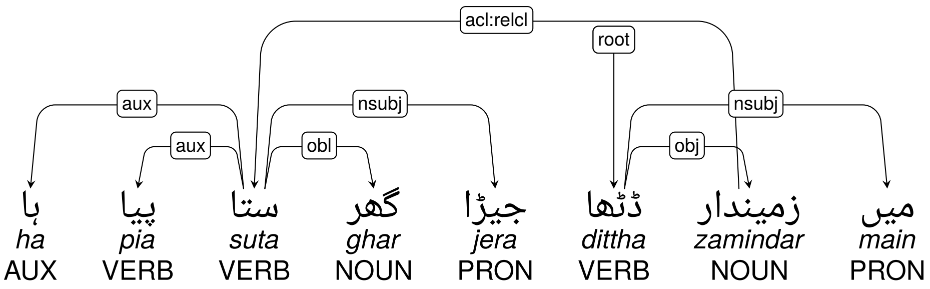


Figure 4: The annotation of the example of an externally headed relative clause in (15).

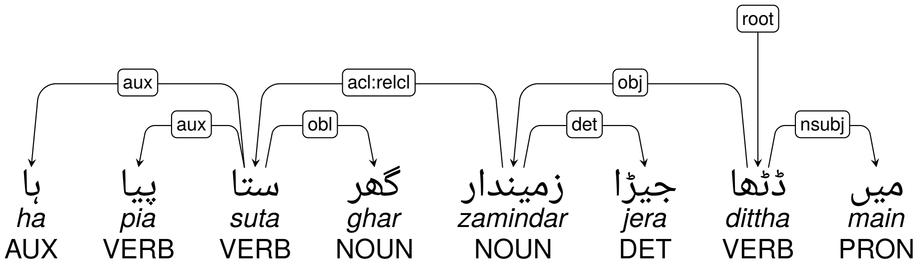


Figure 5: The annotation of the example of an internally headed relative clause in (16).

6.3. Relative Clauses

In the Saraiki treebank, we found both finite and non-finite relative clauses. According to Bashir and Conners (2019), both types of clauses are used freely in Saraiki. While Saraiki uses externally headed relative clauses, it also uses internally headed and correlative forms. Saraiki uses جیڑا (*jera* 'that, which') as a relativizer, which agrees with its head noun in number, gender, and case. These types of constructions are also available in Urdu (Ehsan and Butt, 2020; Bhat and Sharma, 2012) and Punjabi (Arora, 2022).

The examples discussed here are part of the sentences created for analyzing specific constructions in Saraiki. We use those examples so that we can focus on the relevant construction without interference from other syntactic phenomena.

Example (15) shows an externally headed relative clause, the annotation is shown in Figure 4. In such cases, جیڑا (*jera* 'which') functions as relative pronoun; here it modifies زمیندار (*zamindar* 'farmer'). We annotate the relative pronoun as `nsubj` of the verb of the relative clause, ستا (*sutta* 'sleep-pst'), which in turn is dependent on the noun in the matrix clause via the `acl:relcl` relation.

(15) ہا پیا ستا گھر جیڑا
ha pia suta ghar jera
AUX PROG sleep-PST house REL.M.SG
ڈٹھا زمیندار میں
dittha zamindar main
see-PST farmer DIR.1.M.SG
"I saw the farmer who was sleeping in the house"

Example (16) shows a version of the sentence with an internally headed relative clause, the annotation is shown in Figure 5. Here, the head noun زمیندار (*zamindar* 'farmer') occurs inside the relative clause, i.e., between the relative pronoun and the object of the relative clause (گھر *ghar* 'house'). Since this means that the relative clause has a relativizer and the noun it refers to, we have decided that the head noun زمیندار (*zamindar* 'farmer-m-sg') serves as the direct object (`obj`) in the matrix clause, and the relativizer serves as its determiner in a `det` relation. Consequently, the verb of the relative clause is dependent on the head noun via a `acl:relcl` relation. This analysis means that we do not consider the head noun to be part of the relative clause, since it provides the only "attachment site" for the relative clause.

(16) ہا پیا ستا گھر زمیندار
ha pia suta ghar zamindar
AUX PROG sleep-PST house farmer
جیڑا ڈٹھا میں
jera dittha main
REL-M-SG see-PST DIR.1.M.SG
"I saw the farmer who was sleeping in the house"

Example (17) shows the same internally headed version, but in a different word order, with a fronted relative clause. The annotation is shown in Figure 6. Based on our current understanding, we

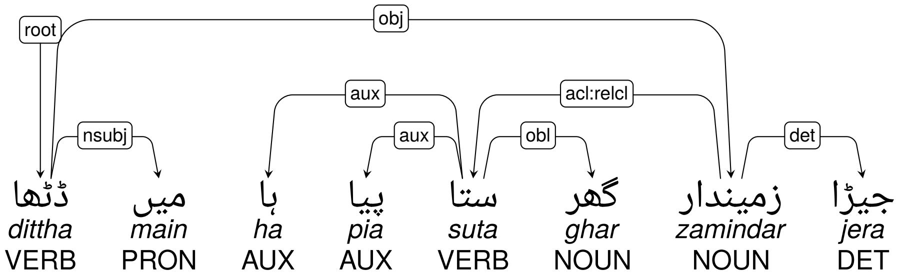


Figure 6: The annotation of the example of an internally headed, fronted relative clause in (17).

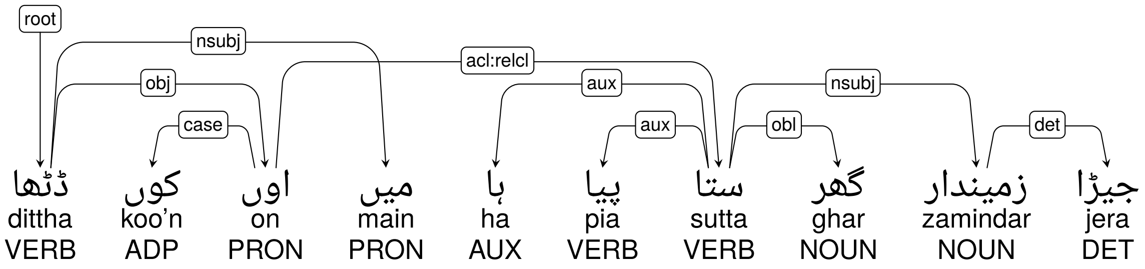


Figure 7: The annotation of the example of a correlative relative clause in (18).

assume that the only difference between all three variants is in information structure.

(17) ڈٹھا میں ہا پیا ستا
dittha main ha pia suta
see-PST DIR.1.M.SG AUX PROG sleep-PST
گھر زمیندار جیڑا
ghar zamindar jera
home farmer REL-M-SG
"I saw the farmer who was sleeping in the house"

Example (18) shows the same sentence, but uses a correlative. The annotation is shown in Figure 7. Correlative relative clauses are a variant of internally headed relative clauses where the relative clause is dependent on, and in an anaphoric relation to, a pronoun in the matrix clause. In example (18), the distal pronoun اوں (*oun* 'that') serves as the correlative. Consequently, we annotate it as the direct object of the matrix clause. The fronted relative clause is dependent on this pronoun. Parallel to the internally headed examples in (16) and (17), we analyze the relativizer as a determiner dependent on the subject of the relative clause.

(18) ڈٹھا اوں کوں میں ہا
dittha koo'n on main ha
see-PST to ACC.3.M.SG DIR.1.M.SG AUX
پیا ستا گھر زمیندار
pia sutta ghar zamindar
PROG sleep-PST house farmer
جیڑا
jera
REL.SG.M
"I saw the farmer who was sleeping in the house"

7. Conclusion and Future Work

We have presented a treebank for Saraiki, annotated using Universal Dependencies. We discussed the textual basis of the treebank and a range of language specific syntactic phenomena. The treebank is work in progress, it currently comprises 587 sentences. We will we will keep extending it and release it once we reach 1 000 sentences.

For future work, we will need to have a closer look at the relative clauses. Additionally, we plan to automatically annotate the morphological features using the Apertium morphological analyzer for Saraiki (Alam et al., 2023). We hope that this treebank will spur deeper investigations of Saraiki as well as the creation of NLP tools for the language. We also plan to train a syntactic parser, and investigate zero-shot techniques to extend our work to other regional languages such as Punjabi (Shahmukhi), Hindko, and Khetrani.

8. Acknowledgements

We would like to thanks Pervaiz Qadir for developing the Saraiki corpus in Mozilla Common Voice and Zahoor Dhareja for giving permission for us to use data from Jhok newspaper for the treebank. We would also like to thanks Daniel Swanson and Daniel Zeman for their help with annotation decisions.

9. Bibliographical References

I. Aduriz, M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. Díaz de Ilarraza, A. Garmendia, and M. Oronoz. 2003. Construction of a Basque dependency treebank. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, pages 201–204, Växjö, Sweden.

Meesum Alam, Alexandra O'Neil, Daniel Swanson, and Francis Tyers. 2023. A finite-state morphological analyzer for Saraiki. In *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL)*, pages 9–13.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 4218–4222, Marseille, France.

Aryaman Arora. 2022. Universal Dependencies for Punjabi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 5705–5711.

Muhammad Nabeel Asghar, Farrukh Javed Saleemi, Sajid Iqbal, Muhammad Umar Chaudhry, Muhammad Yasir, Sibghat Ullah Bazai, and Muhammad Qasim Khan. 2021. A novel parts of speech (POS) tagset for morphological, syntactic and lexical annotations of Saraiki language. *Journal of Applied and Emerging Sciences*, 11(1):pp–77.

András Bárány. 2012. Hungarian conjugations and differential object marking. In *Proceedings of the First Central European Conference in Linguistics for Postgraduate Students*, pages 3–25.

Elena Bashir and Thomas J Conners. 2019. *A Descriptive Grammar of Hindko, Panjabi, and Saraiki*, volume 4. Walter de Gruyter.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, pages 987–998, New Orleans, LA.

Riyaz Ahmad Bhat and Dipti Misra Sharma. 2012. Dependency treebank of Urdu and its evaluation. In *Proceedings of the Sixth Linguistic Annotation Workshop (LAW)*, pages 157–165.

Marie-Catherine de Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Zahoor Dhareja. 2017–2022. Jhok Multan. (Daily newspaper in Pakistan).

Robert MW Dixon. 1994. *Ergativity*. Cambridge University Press.

Toqeer Ehsan and Miriam Butt. 2020. Dependency parsing for Urdu: Resources, conversions and learning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 5202–5207, Marseille, France.

Mandeep Singh Gill, Gurpreet Singh Lehal, and Shiv Sharma Joshi. 2009. Part of speech tagging for grammar checking of Punjabi. *Linguistic Journal*, 4(1):6–21.

Lewis Glinert. 2004. *The Grammar of Modern Hebrew*. Cambridge University Press.

Sarah Gul, Musarrat Azher, and Sana Nawaz. 2021. Development of saraiki wordnet by mapping of word senses: A corpus-based approach. *Linguistics and Literature Review*, 7(2):46–66.

Andrew Hardie. 2003. Developing a tagset for automated part-of-speech tagging in Urdu. In *Corpus Linguistics*.

Andrew John Jukes. 2019. *Dictionary of the Jatki or Western Panjábi language*. Routledge.

de Marneffe M.-C. Ginter F. Hajič J. Manning C. D. Pyysalo S. Schuster S. Tyers F. M. Nivre, J. and D. Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*, Istanbul, Turkey.

Vinit Ravishankar. 2017. A universal dependencies treebank for Marathi. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 190–200.

Christopher Shackle. 1976. *The Siraiki Language of Central Pakistan: A Reference Grammar*. School of Oriental and African Studies, Univ. of London.

Nasir Abbas Syed and Ghulam Raza. 2019. Pronominal suffixes and clitics in Saraiki. *Pakistan Journal of Languages and Translation Studies*, (1):148–173.

Juhi Tandon, Himani Chaudhry, Riyaz Ahmad Bhat, and Dipti Misra Sharma. 2016. Conversion from Paninian karakas to Universal Dependencies for Hindi dependency treebank. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X)*, pages 141–150.

Francis M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2017. UD annotatrix: An annotation tool for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 10–17, Prague, Czech Republic.

Domain-Weighted Batch Sampling for Neural Dependency Parsing

Jacob Striebel, Daniel Dakota, Sandra Kübler
Indiana University
{jstrieb,ddakota,skuebler}@indiana.edu

Abstract

In neural dependency parsing, as well as in the broader field of NLP, domain adaptation remains a challenging problem. When adapting a parser to a target domain, there is a fundamental tension between the need to make use of out-of-domain data and the need to ensure that syntactic characteristic of the target domain are learned. In this work we explore a way to balance these two competing concerns, namely using *domain-weighted batch sampling*, which allows us to use all available training data, while controlling the probability of sampling in- and out-of-domain data when constructing training batches. We conduct experiments using ten natural language domains and find that domain-weighted batch sampling yields substantial performance improvements in all ten domains compared to a baseline of conventional randomized batch sampling.

Keywords: neural dependency parsing, domain adaptation, batch optimization

1. Introduction

Dependency parsing, like many other machine learning problems, is sensitive to domain shifts between training and test data sets (Gildea, 2001; Petrov and Klein, 2007). To combat the negative effects of domain shifts when training a parser, several domain adaptation techniques have been studied (e.g., Rosa and Žabokrtský, 2015), although their effectiveness is often limited (e.g., Dredze et al., 2007).

A major factor that determines the success of domain adaptation methods is the amount of training data that is available in the adaptation-target domain (e.g., Daumé III, 2007; Dredze et al., 2007). To overcome the frequent problem of scarcity of target-domain training data, common techniques in parsing focus on selecting optimal source data points to boost performance in the target domain (Plank and van Noord, 2011; McDonald et al., 2011; Mukherjee and Kübler, 2017), with both delexicalized (Rosa and Žabokrtský, 2015) and lexicalized (Falenska and Çetinoğlu, 2017) similarity metrics showing improved data point selection.

Furthermore, to more effectively use all available source- and target-domain data, discrepancies in sizes between data sources have been handled using loss weighting on the different data sources (Dakota et al., 2021), allowing for noise reduction and improved information sharing.

Other approaches for encoding more domain-related information into a parser are to create data- or task-specific embeddings (Stymne et al., 2018; Li et al., 2019, 2020), which yield performance gains across languages and domains. While the further inclusion of language models into parsing architectures noticeably reduces performance gaps across domains, it still cannot fully overcome syntactic differences (Joshi et al., 2018; Fried et al., 2019; Yang et al., 2022). The situation is further complicated by the fact that the source and target domains may be different from those of the language model (Dakota, 2021).

We focus on a setting in which we have access to a small amount of annotated data from the target domain. In order to address the size difference between the data available for the target domain and other domains, we investigate a method that allows the use of all available source and target data during training, thus maximizing the available signal. More specifically, we use *domain-weighted batch sampling* (DWBS) to train a domain-expert neural dependency parser as an alternative to the conventional approach of *randomized batch sampling* (RBS).

Since we use some target domain data for training in our experiments, existing naming conventions are not easily usable. For this reason, we call data from the target domain *in-domain data* and data from all other domains *out-of-domain data* (i.e., any domain that is not the adaptation-target domain); we also use *source data* as a synonym for out-of-domain data. Note that our sampling strategy can also be used when we do not have any in-domain data but can determine the most similar domain among the out-of-domain data.

Our experiments are designed to answer the following two questions:

1. Can we improve parser performance, given a training data imbalance between in-domain and out-of domain data, by replacing the standard batch sampling approach (i.e., RBS) with DWBS, which uses all available training data but favors training sentences drawn from the target evaluation domain?

2. Does DWBS yield faster training times than RBS? In other words, does DWBS reduce the number of sample sentences that a parser must observe before dev loss stops decreasing?

2. Domain-Weighted Batch Sampling

2.1. Batch Sampling

When training a neural network, there are several approaches that can be taken to creating batches, and the chosen approach will impact how a network converges, memory requirements, and possible performance among other effects on the model.

The simplest way of creating a batch is to select training samples in the order in which they appear in the training data file, which is called *sequential batch sampling* (SBS). However, this strategy may not be optimal since it repeatedly exposes the network to the same sequence of examples and thus may cause the network to indirectly learn specific batch characteristics that are not representative of the task as a whole (Chollet, 2018), which can result in catastrophic forgetting (French, 1999; Dachapally and Jones, 2018). Consequently, it is more common to create randomized permutations of the training data at the beginning of every epoch, which is called *randomized batch sampling* (RBS).

2.2. Domain-Weighted Batch Sampling

To leverage in-domain and all out-of-domain data, we extend RBS to *domain-weighted batch sampling* (DWBS). This allows for better inclusion of multi-source out-of-domain data, while still permitting the target domain to maintain higher influence on optimization.

To perform DWBS, before training begins the training data set is partitioned into disjoint *in-domain* and *out-of-domain* subsets. For each epoch, random permutations of the in-domain and out-of-domain subsets are separately generated. Each batch is then constructed by drawing sentences (without replacement) from the two permutations until the batch size is reached. We use the hyperparameter μ to define the probability of choosing the next sentence from the in-domain permutation. For example, if μ is equal to 0.45, there is a 45% chance of drawing the next sentence from the target (in-domain) permutation and 55% of drawing from the source (out-of-domain) permutation.

During an epoch, eventually we will attempt to draw from a permutation in which no sentences remain, at which point the current partially constructed batch is discarded and the current epoch is complete. A side-effect of the DWBS procedure is that different epochs may have different durations in terms of number of batches.

Hyperparameter	Value
Optimizer	Adamw
β_1, β_2	0.9, 0.99
Correction bias	False
Learning rate	0.0001
Weight decay	0.01
Gradient normalization	1
LR scheduler	Slanted triangular
Cut fraction	0.2
Decay factor	0.38
Discriminative fine tuning	True
Gradual unfreezing	True
Batch size	32
Patience batches	200
Max steps	153,600
Embeddings	bert-base-cased
Embeddings dim	768

Table 1: Hyperparameters

3. Methodology

3.1. Data

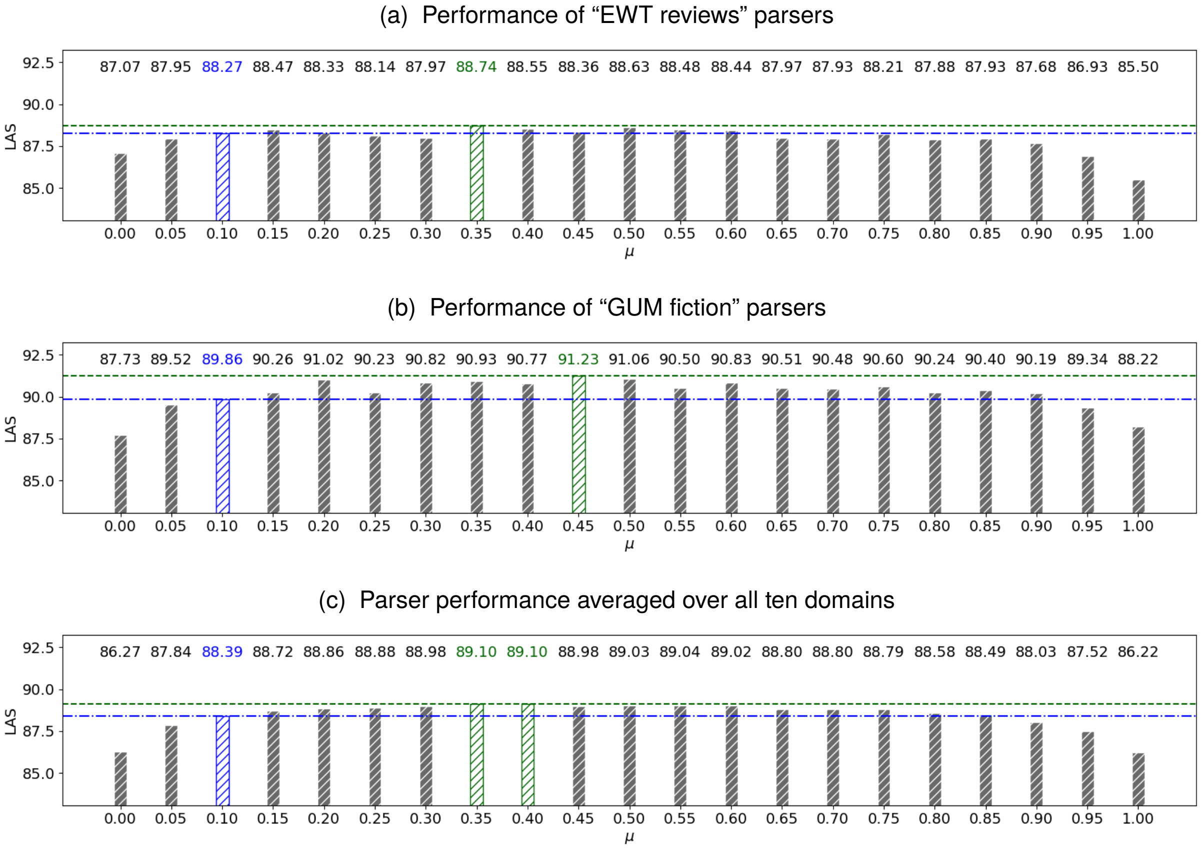
We use Universal Dependency treebanks version 2.12 (Nivre et al., 2020; de Marneffe et al., 2021), more specifically the English Web Treebank (EWT; Bies et al., 2012) and the Georgetown University Multilayer Corpus (GUM; Zeldes, 2017). EWT consists of five domains, and GUM consists of eleven domains.

From the sixteen domains of EWT and GUM, we select only the ten domains that each have a minimum of 1 000 sentences, to limit negative effects during training due to different data sizes across domains. This includes all five of the EWT domains: answers, email, newsgroup, reviews, weblogs; and five from GUM: conversation, fiction, interviews, vlog and whow. We then randomly sub-sample only 1 000 sentences from each domain to create a balanced data set.

All of our experiments use ten-fold cross validation, where, for each fold, each domain is split into 800 train, 100 dev, and 100 test sentences. Consequently, when training each domain-expert parser, there are a total of 8 000 train sentences (800 in-domain and 7 200 out-of-domain), and 100 dev and 100 test sentences (all of these in-domain).

3.2. Parser

We use the deep biaffine attention neural dependency parser (Dozat and Manning, 2017) in the implementation by van der Goot et al. (2021b), which we have modified to allow for DWBS. When training the parser, we use the default hyperparameters provided by van der Goot et al., with the only exception being that we specify early-stopping patience


(a) Performance of “EWT reviews” parsers

(b) Performance of “GUM fiction” parsers

(c) Parser performance averaged over all ten domains

Figure 1: Performance of the DWBS-trained domain-expert parsers on “EWT reviews” (a), “GUM fiction” (b), and averaged over all ten domains (c). X-axis: domain-weight hyperparameter μ; y-axis: parser performance in LAS. Because in our experimental setup we use use ten domains of equal size, whenever μ = 0.10, DWBS is equivalent to conventional RBS; therefore, in each chart we highlight the baseline RBS-trained parser in blue, and we highlight the best performing DWBS-trained parser(s) in green.

in terms of batches rather than epochs, because, when DWBS is enabled, epoch duration varies with μ and it is also subject to random variation (see Section 2.2). Batch size, on the other hand, is a fixed hyperparameter. All hyperparameters are reported in Table 1.

For each domain, and for each of the ten data folds, we use the dev sentences to determine when to stop training, and we then use the test sentences to evaluate. We evaluate using the scorer from the CoNLL 2018 shared task (Zeman et al., 2018).

4. Results

In order to evaluate the effectiveness of DWBS, we perform experiments in which we compare a baseline model trained using conventional RBS against domain-expert parsers trained using DWBS. For each domain, we train domain-expert parsers, with the domain-weight hyperparameter μ ranging from 0.00 to 1.00 (inclusive), with a step size of 0.05. Remember that μ = 0.00 means that each batch will be sampled exclusively from the *out-of-domain* par-

TB	Domain	μ	LAS R	LAS DW
EWT	Answers	0.35	86.78	87.56
	Email	0.35	86.70	**88.00**
	Newsgr.	0.40	88.64	89.44
	Reviews	0.35	88.27	88.74
	Weblog	0.25	89.52	**90.56**
GUM	Convers.	0.35	85.41	**86.64**
	Fiction	0.45	89.86	**91.23**
	Interv.	0.50	88.08	**89.14**
	Vlog	0.60	87.74	88.57
	Whow	0.35	90.46	91.11

Table 2: Performance in LAS per domain, comparing the baseline parser (trained using RBS) to the highest-LAS-producing domain-expert parser (trained using DWBS). LAS R: baseline parser trained using RBS; LAS DW: highest-LAS-producing domain-expert parser trained using DWBS; μ: setting resulting in the highest LAS for the given domain. Improvements of more than 1.00 LAS are bolded.

Treeb.	Domain	μ	RBS NSC	DWBS NSC	ΔNSC
EWT	Answers	0.35	40.40	40.88	0.48
	Email	0.35	39.84	40.00	0.16
	Newsgroup	0.40	45.60	45.44	-0.16
	Reviews	0.35	40.96	41.36	0.40
	Weblog	0.25	47.04	48.00	0.96
GUM	Conversation	0.35	45.52	41.20	-4.32
	Fiction	0.45	40.56	42.96	2.40
	Interview	0.50	45.20	42.00	-3.20
	Vlog	0.60	48.16	42.96	-5.20
	Whow	0.35	40.40	40.24	-0.16

Table 3: Training duration per domain measured in number of thousands of samples until model convergence, comparing the baseline parser to the highest-LAS-producing domain-expert parser. NSC: number of thousands of training samples until model convergence; RBS NSC: NSC for the baseline parser trained using RBS; DWBS NSC: NSC for the highest-LAS-producing domain-expert parser trained using DWBS; μ: setting yielding the best (in terms of LAS) domain-expert parser for the given domain.

tition of the training data set, while μ = 1.00 means that training samples will only be drawn from the *in-domain* partition. Because our training data set is composed of ten domains of equal size, DWBS for μ = 0.10 is equivalent to conventional RBS.

4.1. Effect on Parsing Accuracy

The DWBS-trained parser outperforms the baseline in all ten domains tested, for some settings of μ. We provide full results for two domains, plus the results averaged over all ten domains, in Figure 1; full results for the remaining domains are supplied in Appendix A. Table 2 summarizes the results by giving the LAS for the highest performing DWBS-trained parser, per domain, and giving the setting for μ that produced the parser.

The domain which benefits least from DWBS, in terms of absolute increase in LAS over the baseline, is EWT reviews, for which the best setting of μ = 0.35 yields an improvement of 0.47 LAS (see Figure 1a); the domain benefiting most is GUM fiction, for which the best setting of μ = 0.45 gives an improvement of 1.37 LAS (see Figure 1b). The average improvement across all ten domains, using each domain's best setting of μ, is 0.95 LAS. As shown in Table 2, five domains experience gains of more than 1.00 LAS.

Overall, the best setting of μ ranges between 0.25 (EWT weblog) and 0.60 (GUM vlog). GUM domains tends to prefer higher values of μ. In other words, those domains profit more from training examples from the same domain, which is an indication that each of those domains is different from all others, either in terms of syntactic structure or annotation.

4.2. Effect on Training Duration

Our hypothesis wrt training times is that the more target-domain sentences that are included in training batches, the faster the parser should converge, since the training sentences should be more consistent and also more similar to the dev data. This hypothesis is supported by findings that alternative batch sampling techniques to RBS which are similarly motivated to DWBS yield significantly faster network training times on several tasks (Loshchilov and Hutter, 2016).

We show the average number of training examples until model convergence for the highest-LAS-producing μ per domain in Table 3. In contrast to the results presented in the previous subsection in which all ten domains show an improvement in LAS, the domains are evenly split on training time reduction with five seeing a reduction and five experiencing an increase. The greatest increase is experienced by the GUM fiction domain, which requires 2400 more sentences than the baseline to achieve parser convergence, while the greatest decrease is experienced by the GUM vlog domain, which shows a decrease of 5200 sentences until convergence. The average change in training samples is a decrease of 864 sentences. The high variability of differences in training duration suggests that DWBS does not reliably reduce the number of samples required to achieve parser convergence. This may suggest that our target domain data do not always have high internal consistency, which is in line with findings by Zeldes and Schneider (2023), who observed considerable differences in cross-domain parsing between EWT and GUM.

Interestingly, four out of the five domains showing decreased training times are GUM domains. Since GUM domains also prefer higher values of μ, this could suggest that sampling more target sentences reduces training time.

5. Conclusion

In this work we investigated the effectiveness of domain-weighted batch sampling (DWBS) when training a neural dependency parser. DWBS is a technique for constructing training batches that can be used in cases when the domain that a parser will be evaluated on is known and there is also training data available in the evaluation domain. We conducted experiments using ten English domains and found that DWBS produced higher performing parsers than RBS in all ten domains. This finding suggests that when the preconditions for performing DWBS are met, it should be preferred to RBS when training a neural dependency parser.

The success of DWBS for neural dependency parsing suggests several directions for future work: In the present experiment while training each model, the domain-weight parameter μ was held constant for the full duration of training. An alternative is to begin training with μ equal to the baseline setting, and then gradually increase μ as training progresses. This will simulate *gradually fine-tuning* the parser in the target domain. A second area of future work is to experiment with methods of automatically classifying domains (e.g., in the style of Mukherjee et al., 2017; Mukherjee and Kübler, 2017), which would allow for the discovery of more syntactically useful domain groupings. Finally, we will investigate the effectiveness of domain embeddings (van der Goot and de Lhoneux, 2021; van der Goot et al., 2021a; Li et al., 2019, 2020), an alternative approach to domain adaptation in dependency parsing that can be combined with domain-weighted batch sampling.

6. Acknowledgments

The authors acknowledge the Indiana University Pervasive Technology Institute for providing supercomputing and storage resources that have contributed to the research results reported within this paper.

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

7. Bibliographical References

Francois Chollet. 2018. *Deep Learning with Python*. Manning, Shelter Island, NY.

Prudhvi Dachapally and Michael Jones. 2018. Catastrophic interference in neural embeddings models. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

Daniel Dakota. 2021. Genres, parsers, and BERT: The interaction between parsers and BERT models in cross-genre constituency parsing in English and Swedish. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 59–71, Online. Association for Computational Linguistics.

Daniel Dakota, Zeeshan Ali Sayyed, and Sandra Kübler. 2021. Bidirectional domain adaptation using weighted multi-task learning. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 93–105, Online. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João Graca, and Fernando Pereira. 2007. Frustratingly hard domain adaptation for dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1051–1055, Prague, Czech Republic.

Agnieszka Falenska and Özlem Çetinoğlu. 2017. Lexicalized vs. delexicalized parsing in low-resource scenarios. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 18–24, Pisa, Italy.

Robert French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3:128–135.

Daniel Fried, Nikita Kitaev, and Dan Klein. 2019. Cross-domain generalization of neural constituency parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 323–330, Florence, Italy.

Daniel Gildea. 2001. Corpus Variation and Parser Performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 167–202, Pittsburgh, PA.

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1190–1199, Melbourne, Australia.

Ying Li, Zhenghua Li, and Min Zhang. 2020. Semi-supervised domain adaptation for dependency parsing via improved contextualized word representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3806–3817, Barcelona, Spain (Online).

Zhenghua Li, Xue Peng, Min Zhang, Rui Wang, and Luo Si. 2019. Semi-supervised domain adaptation for dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2386–2395, Florence, Italy.

Ilya Loshchilov and Frank Hutter. 2016. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK.

Atreyee Mukherjee and Sandra Kübler. 2017. Similarity based genre identification for POS tagging experts & dependency parsing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 519–526, Varna, Bulgaria. INCOMA Ltd.

Atreyee Mukherjee, Sandra Kübler, and Matthias Scheutz. 2017. Creating POS tagging and dependency parsing experts via topic modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 347–355, Valencia, Spain. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4034–4043, Marseille, France.

Slav Petrov and Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 404–411, Rochester, NY.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA.

Rudolf Rosa and Zdeněk Žabokrtský. 2015. KLcpos3 - a language similarity measure for delexicalized parser transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 243–249, Beijing, China.

Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre. 2018. Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia. Association for Computational Linguistics.

Rob van der Goot and Miryam de Lhoneux. 2021. Parsing with pretrained language models, multiple datasets, and dataset embeddings. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 96–104, Sofia, Bulgaria. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, and Barbara Plank. 2021a. On the effectiveness of dataset embeddings in mono-lingual,multi-lingual and zero-shot conditions. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 183–194, Kyiv, Ukraine. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021b. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics:*

System Demonstrations, pages 176–197, Online. Association for Computational Linguistics.

Sen Yang, Leyang Cui, Ruoxi Ning, Di Wu, and Yue Zhang. 2022. Challenges to open-domain constituency parsing. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 112–127, Dublin, Ireland. Association for Computational Linguistics.

Amir Zeldes and Nathan Schneider. 2023. Are UD treebanks getting more consistent? a report card for English UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 58–64, Washington, D.C.

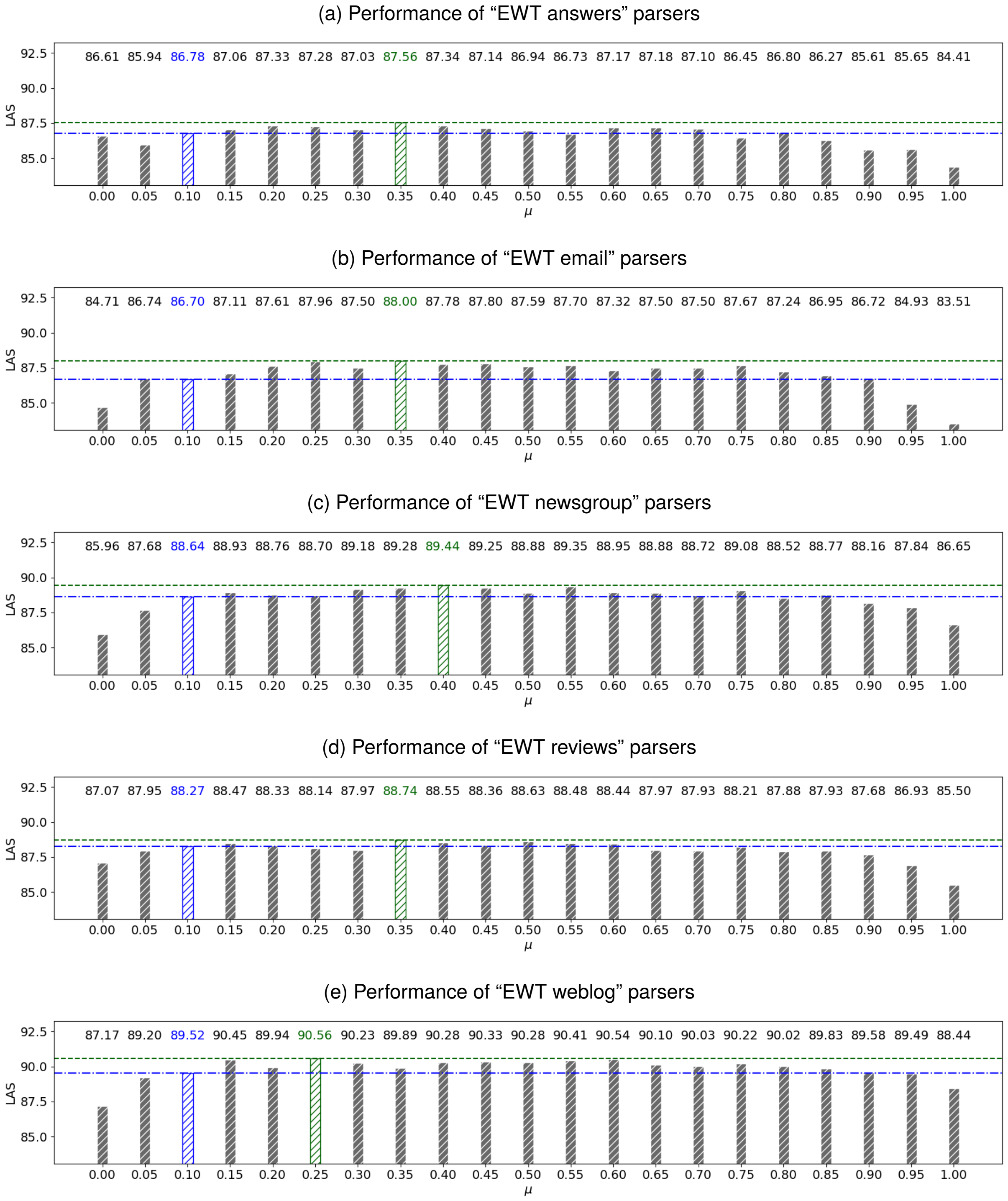
Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium.

8. Language Resource References

Bies, Ann and Mott, Justin and Warner, Colin and Kulick, Seth. 2012. *English Web Treebank LDC2012T13*. Linguistic Data Consortium.

Amir Zeldes. 2017. *Georgetown Multilayer Corpus (GUM)*. Georgetown University Corpus Linguistics Lab.

A. Complete Parsing Results



(a) Performance of “EWT answers” parsers

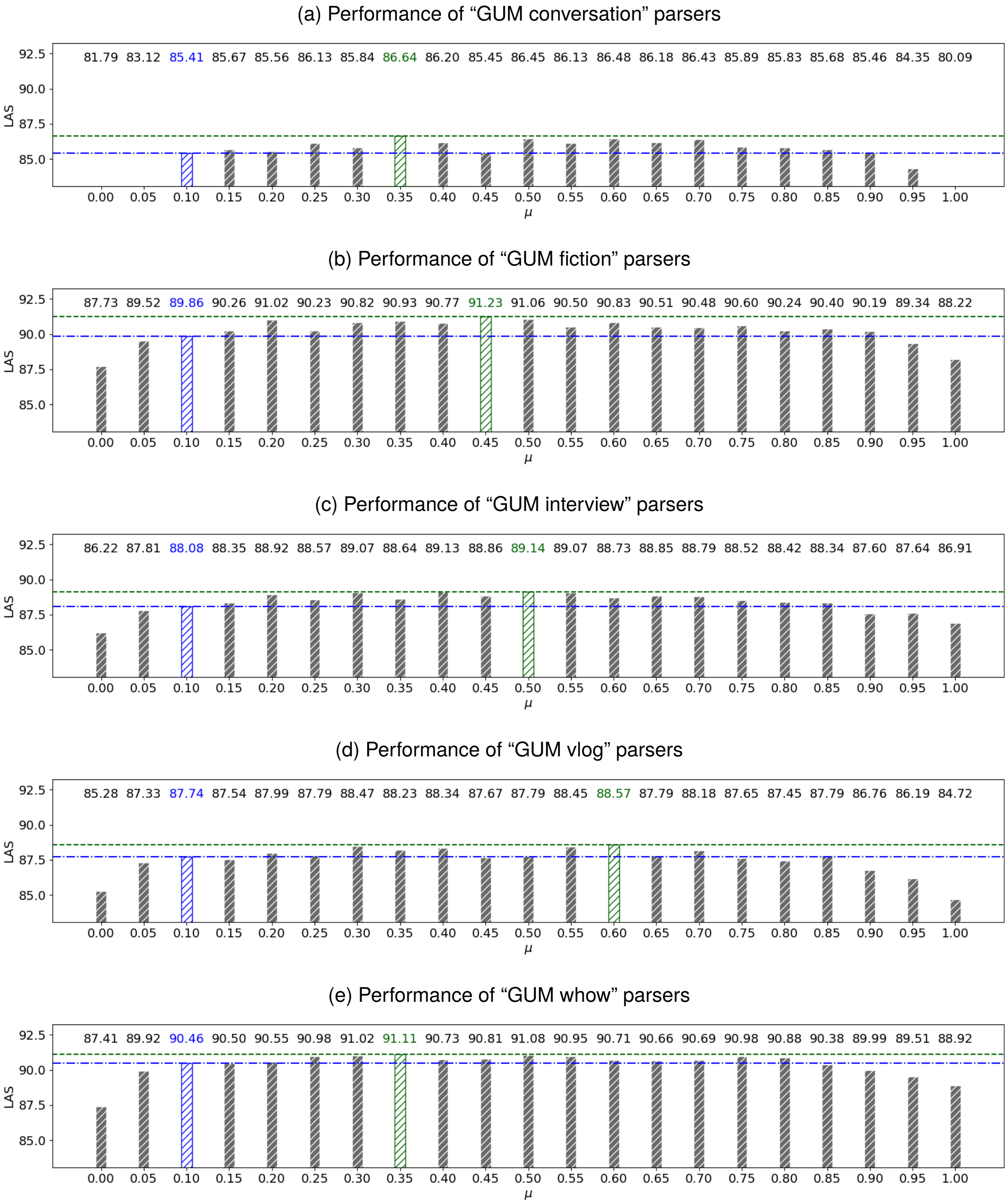
(b) Performance of “EWT email” parsers

(c) Performance of “EWT newsgroup” parsers

(d) Performance of “EWT reviews” parsers

(e) Performance of “EWT weblog” parsers

Figure 2: Parser performance in the five **English Web Treebank** domains. X-axis: domain-weight hyperparameter μ; y-axis: parser performance (LAS). Baseline RBS-trained parser in blue, and best performing DWBS-trained parser in green.



(a) Performance of “GUM conversation” parsers

(b) Performance of “GUM fiction” parsers

(c) Performance of “GUM interview” parsers

(d) Performance of “GUM vlog” parsers

(e) Performance of “GUM whow” parsers

Figure 3: Parser performance in the five **Georgetown University Multilayer Corpus** domains. X-axis: domain-weight hyperparameter μ; y-axis: parser performance (LAS). Baseline RBS-trained parser in blue, and best performing DWBS-trained parser in green.

Strategies for the Annotation of Pronominalised Locatives in Turkic Universal Dependency Treebanks

Jonathan Washington[1], **Çağrı Çöltekin**[2], **Furkan Akkurt**[3], **Bermet Chontaeva**[2], **Soudabeh Eslami**[2], **Gulnura Jumalieva**[4], **Aida Kasieva**[4], **Aslı Kuzgun**, **Büşra Marşan**[5], **Chihiro Taguchi**[6]

[1]Swarthmore College, [2]University of Tübingen, [3]Boğaziçi University,
[4]Kyrgyz-Turkish Manas University, [5]Stanford University, [6]University of Notre Dame,
jwashin1@swarthmore.edu,cagri.coeltekin@uni-tuebingen.de,furkan.akkurt@bogazici.edu.tr,
{bermet.chontaeva,soudabeh.eslami}@student.uni-tuebingen.de,{gulnur.jumalieva,
aida.kasieva}@manas.edu.kg,kuzgunasli@gmail.com,
busra@stanford.edu,ctaguchi@nd.edu

Abstract

As part of our efforts to develop unified Universal Dependencies (UD) guidelines for Turkic languages, we evaluate multiple approaches to a difficult morphosyntactic phenomenon, pronominal locative expressions formed by a suffix *-ki*. These forms result in multiple syntactic words, with potentially conflicting morphological features, and participating in different dependency relations. We describe multiple approaches to the problem in current (and upcoming) Turkic UD treebanks, and show that none of them offers a solution that satisfies a number of constraints we consider (including constraints imposed by UD guidelines). This calls for a compromise with the 'least damage' that should be adopted by most, if not all, Turkic treebanks. Our discussion of the phenomenon and various annotation approaches may also help treebanking efforts for other languages or language families with similar constructions.

Keywords: Turkic languages, Universal Dependencies, treebanks

1. Introduction

As the number of treebanks for a single language or a language family in the Universal Dependencies (UD) repository[1] grows, consistent annotations become a concern (Gamba and Zeman, 2023a,b; Zeldes and Schneider, 2023). We report on one issue that is part of ongoing efforts to unify Universal Dependencies (UD) treebanks for Turkic languages, currently numbering at 16 in 8 different UD languages. Issues regarding the consistency of UD annotation of Turkic languages have been reported in earlier studies (Tyers et al., 2017; Türk et al., 2019; Çöltekin et al., 2022), with the main consensus being the need for more unified and consistent annotations across treebanks.

In this paper, we examine one selected issue in depth—namely, that of *-ki*, which attaches to nouns in the genitive and locative case. With locative nouns, it forms either attributive expressions or pronominals, while with genitive nouns, the result is always a pronominal expression.[2] As explained in detail in §2, how to appropriately annotate these pronominal forms is unclear and problematic with the present UD guidelines. As a result, the current Turkic treebanks adopt different approaches to annotating this construction. Divergence also exists within different treebanks of the same language.

We believe that the discussion of this linguistic phenomenon is likely to increase the consistency of current treebanks, help researchers creating new treebanks for Turkic languages (and others facing similar issues), and may result in improvements to the general UD guidelines by highlighting issues that are not well addressed in the current guidelines.

In this paper, we provide background information on the issue of pronominalised locatives (§2), discuss in depth several possibilities for the annotation of pronominalised locatives in Turkic languages (§3), summarise these approaches (§4), and conclude (§5). While a recommendation for a preferred approach is not put forth, a potential compromise is identified.

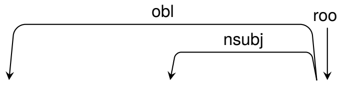
2. The issue of pronominalised locatives

In Turkic languages, locative forms of nominals (e.g., nouns, pronouns, and proper nouns) function as a locative adjunct/modifier to the head of an embedded or root clause, as in Figure 1.

Locatives cannot modify nouns on their own. One common strategy to use locatives attributively as a modifier to a noun is with the addition of the

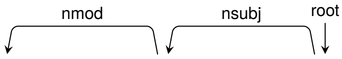
[1] See Appendix A for information on current and upcoming Turkic UD treebanks.

[2] Here, we only focus on the more varied, locative version. The outcome of the present discussion is likely to inform the issue of the annotation of genitives as well.



Turkish:	*oda-da*	*çocuk-lar*	*uyu-du-lar*
Azerbaijani:	*otaq-da*	*uşaq-lar*	*yuxla-dı-lar*
Kyrgyz:	*бөлмө-дө*	*бал-дар*	*укта-ды*
Tatar:	*бүлмә-дә*	*бала-лар*	*йокла-ды*
	room-LOC	child-PL	sleep-PST(-PL)
Lemma:	oda/otaq	çocuk/uşaq	uyu/yuxla
	/бөлмө/бүлмә	/бала/бала	/укта/йокла
POS:	NOUN	NOUN	VERB
Case:	Loc	Nom	-

Figure 1: A sentence containing an attributive locative; English translation: "Children slept in the room."



Turkish:	*oda-da-ki*	*çocuk-lar*	*uyu-du-lar*
Azerbaijani:	*otaq-da-kı*	*uşaq-lar*	*yuxla-dı-lar*
Kyrgyz:	*бөлмө-дө-гү*	*бал-дар*	*укта-ды*
Tatar:	*бүлмә-дә-ге*	*бала-лар*	*йокла-ды*
	room-LOC-ATTR	child-PL	sleep-PST(-PL)
Lemma:	oda/otaq	çocuk/uşaq	uyu/yuxla
	/бөлмө/бүлмә	/бала/бала	/укта/йокла
POS:	NOUN	NOUN	VERB
Case:	Loc	Nom	-

Figure 2: A sentence containing an attributive locative; English translation: "The children in the room fell asleep."

morpheme *-ki*,[3] as in Figure 2.[4]

When a locative is used attributively in this way, we opt to annotate it as `nmod` or `nmod:loc`,[5] since it is a nominal dependent (with a noun POS and lemma) of a nominal, just as in the semantically equivalent English sentence. A disadvantage of this approach is that the `Case` feature remains `Loc` and the *-ki* morpheme is not treated separately. However, the structure is recoverable, as these constructions are unique (in each language where it occurs) as the only time a locative `nmod` dependent is found.

As with other attributive expressions in Turkic languages—including adjectives per Krejci and Glass (2015) and verbal adjectives per Washington et al. (2022)—these attributive locative expressions may be used nominally, as a sort of pronominal.[6] We consider this a form of syntactic derivation.[7] For example, the sentence in Figure 2 may be expressed without the noun head of the *-ki* bearing form, with any morphology normally found there being found on the dependent *-ki* bearing form instead, as in Figure 3.

Turkish:	*oda-da-ki-ler*	*uyu-du-lar*
Azerbaijani:	*otaq-da-kı-lar*	*yuxla-dı-lar*
Kyrgyz:	*бөлмө-дө-гү-лөр*	*укта-ды*
Tatar:	*бүлмә-дә-ге-ләр*	*йокла-ды*
	room-LOC-ATTR-PL	sleep-PST(-PL)
Lemma:	oda/otaq	uyu/yuxla
	/бөлмө/бүлмә	/укта/йокла
POS:	NOUN	VERB

Figure 3: A sentence containing a pronominalised locative; English translation "The ones in the room slept."

The resulting pronominal is formed from one noun (in this case, the room), and refers to another referent (such as the children, in this case). Several problems arise from this type of construction since there are two semantic referents (in this case, the room and the ones sleeping there) represented by a single token. Each referent has its own case, number, possessor, and other nominal features expressed through the morphology. While the locative referent still has an `nmod` relation to the other referent and contributes the `Lemma` on which the form is built, it is the other referent that has external relations: in this example, the pronominal is `nsubj` of the `root`. Conversely, the noun would be the head of any adjectival or other dependents. For example, if we add *büyük* 'big' to the Turkish sentence, *büyük odadakiler* has two hypothetical dependency interpretations: (1) 'the ones in the big room' (*büyük* 'big' modifying *oda* 'room'), which is the correct interpretation, and (2) 'the big ones in the room' (*büyük* modifying *odadakiler* 'the ones in the room') is not a possible interpretation. Any solution to annotation that considers the word as a single syntactic unit cannot

[3] In many Turkic languages this has phonologically reduced, e.g. to *-kI* (Azerbaijani) or *-GI* (Kyrgyz, Tatar).

[4] Turkish, Azerbaijani, Kyrgyz, and Tatar are presented as they are the Turkic languages whose UD annotation is currently being considered by the authors.

[5] These two approaches are both acceptable in our opinion, although the latter is more specific and may make identification of this construction easier, for example in an information extraction task.

[6] By 'pronominal', we mean that the resulting form is not a nominal but stands in for one. For example, in Turkish ***büyükleri*** *beğendim* 'I liked **the big ones**', the derived form of the adjective *büyük* 'big' has nominal morphology and refers to an unmentioned nominal. See Göksel and Kerslake (2005, p.246) for a detailed discussion.

[7] I.e., this is a productive process that occurs in the syntax. This is not to be confused with lexical derivation, which is a historical and often not fully productive process and is usually opaque to syntax. Multiple opinions exist as to the specific mechanism by which this pronominalisation operates: through ellipsis of a nominal head, through a null-headed DP, through syntactic transformations, or otherwise.

distinguish these syntactic dependencies. Moreover, such an annotation strategy implies the latter structure, where *büyük* modifies the entire token *odadakiler*.

In an ideal solution to annotation, all morphological and syntactic information about the two participants would be recoverable.

To further complicate matters, the *-ki* morpheme can be attached to the same word multiple times. Although forms with multiple *-ki* morphemes can be difficult to interpret and rare in real-world usage, there is no principled limit for the number of *-ki* morphemes that can be attached to a noun. For example, to refer to 'glasses in the cupboard in the room', we could use the Turkish expression *oda-da-**ki**-nde-**ki**-ler* 'the ones in the one in the room'. Except cognitive load, there is nothing stopping a speaker to add another *-de-ki* to refer to the drinks inside the glasses. Although we will limit our discussion to forms with a single *-ki* morpheme, the ideal solution should also work well for words with multiple occurrences of the morpheme.

In summary, considering the pronominal forms created with the morpheme *-ki* as single syntactic words results in two major issues (see Çöltekin, 2016, for an earlier discussion):

- It violates the *lexical integrity principle* (Haspelmath and Sims, 2010, p.203) since the syntactic dependencies refer to parts of words.

- It also results in conflicting morphological features. For example, in the example in Figure 3, 'room' is singular, while the resulting pronominal refers to multiple people in the room.

The following sections discuss various ways we see as possible approaches to annotating these nominalised constructions in UD.

3. Possible Approaches

Here we demonstrate four possible approaches to the annotation of pronominalised locative forms and discuss advantages and disadvantages of each: keeping a single token (3.1), using layered features (3.2), splitting the token before *-ki* (3.3), and splitting the token after *-ki* (3.4).

We will use the Turkish sentence *Bardak dolabında**ki**lerim düştüler* '**The one**s of mine on the cup cabinet fell' to illustrate how different approaches handle these forms.

The pronominal in this sentence refers to a group of items, e.g., glasses, papers, etc. This example was chosen because there are different number, case, and possession features morphologically indicated for each of the two referents of the pronominalised locative token (the referent of the noun it is formed around and the referent of the pronominal it comprises). An alternative version of this sentence with an independent noun modified by a *-ki* bearing form is provided with annotation in Figure 4 for reference.

3.1. No segmentation

The first option is to have **no segmentation** of the word *dolabındakilerim* 'the ones of mine on its cabinet', as presented in Figure 5.

The advantage of this choice is practical: subword segmentation is a non-trivial task, and avoiding it will help make automated segmentation more precise, especially in low-resource settings. On the other hand, it is not clear what values to assign to the `Number`, `Person[psor]`, or `Person` categories, since the values for both referents of the token *dolabındakilerim* are present: the noun is singular, locative, and has a third-person possessor, while the resulting pronominal is plural, nominative, and has a first-person (plural) possessor.[8] This choice additionally fails to capture several aspects of the dependencies in this sentence:

- that there are two referents of the form: a noun and a pronominal;
- that there is a relationship between the form's two referents;
- that the first noun token in the sentence is a possessor `nmod` of the form's first referent (the noun) and not the second (the pronominal); and
- that the second referent of the form (the pronominal) and not the first (the noun) is the `nsubj` of the `root`.

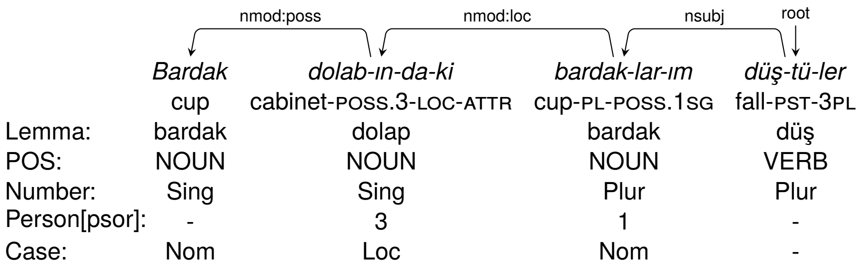
Current treebanks employing a no-segmentation approach in Turkish[9] assume an analysis of elision and use the concept of promotion (whereby a normally dependent function word is 'promoted' to the syntactic function that an elided head would normally have)[10] to annotate dependencies. In our example *oda-da-ki-ler* 'the ones in the room', this approach considers the head word *çocuk-lar* 'children' to be elided.[11] Hence, its dependent *odadaki* is promoted to

[8] All other combinations are also possible in other contexts; for example, *dolaplarındakilerim*, *dolaplarındakim*, or *dolabındakim*.

[9] I.e., Penn (Cesur et al., 2023a), KeNet (Kuzgun et al., 2023b), FrameNet (Cesur et al., 2023b), Tourism (Kuzgun et al., 2023a), Atis (Köse and Yıldız, 2023).

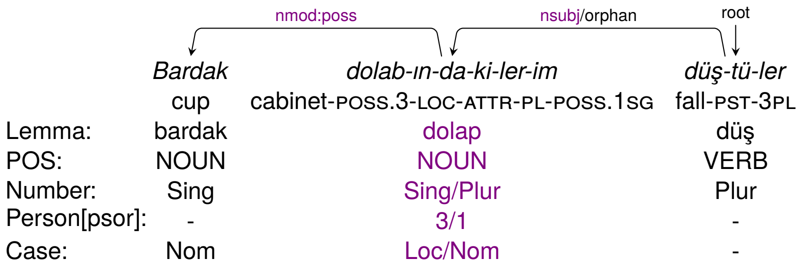
[10] Per `https://universaldependencies.org/u/overview/syntax.html`.

[11] Unlike the English translation where the pronoun *one* still occupies the head of the construction.



	Bardak	*dolab-ın-da-ki*	*bardak-lar-ım*	*düş-tü-ler*
	cup	cabinet-POSS.3-LOC-ATTR	cup-PL-POSS.1SG	fall-PST-3PL
Lemma:	bardak	dolap	bardak	düş
POS:	NOUN	NOUN	NOUN	VERB
Number:	Sing	Sing	Plur	Plur
Person[psor]:	-	3	1	-
Case:	Nom	Loc	Nom	-

Figure 4: Analysis of a sentence comparable to the reference sentence but with a full noun phrase.



	Bardak	*dolab-ın-da-ki-ler-im*	*düş-tü-ler*
	cup	cabinet-POSS.3-LOC-ATTR-PL-POSS.1SG	fall-PST-3PL
Lemma:	bardak	dolap	düş
POS:	NOUN	NOUN	VERB
Number:	Sing	Sing/Plur	Plur
Person[psor]:	-	3/1	-
Case:	Nom	Loc/Nom	-

Figure 5: Analyzing *-ki* with no segmentation.

nsubj. According to this approach as taken in these treebanks, the *-ki* bearing form in the example in Figure 5 is an nsubj dependent of the verb.

Using a Case=Loc feature (as opposed to Case=Nom) with, for example, an nsubj dependent could clarify that this pronominal has some special status. However, a naïve downstream interpretation may understand this to be, in this example, an oblique (locative-marked) subject as opposed to a pronominal locative, especially given that the lemma is that of the attributive word (here, *dolap* 'cabinet') as opposed to the referent to which the morphology and head dependency refer (here, the pronominal referring to e.g., *bardak* 'cup'). Therefore, one option is to use the orphan tag when the *-ki* word is pronominal, shown as an option in Figure 5. The orphan relation is traditionally used in cases of head ellipsis where there is a remnant nominal that must attach to a head that it would not normally attach to. This approach solves the issue with misleading annotations; however, the orphan analysis is not informative. Furthermore, the issues with multiple Number, Person[psor], and Case features that need to be assigned to the form *odadakiler* remain.

Another option is to introduce a new case feature for attributive and pronominal locative, such as AttrLoc. In pronominal uses, as shown in Figure 6, it would then be clear that this structure is not, for example, an oblique subject form of the lemma, but a pronominalised form of an attributive locative formed around the lemma. This at first appears to solve the problem having multiple case features, but it is still not clear how to annotate the second case feature (which can be any of the cases available in a given Turkic language). The problems of multiple number features and possessor person features also remain.

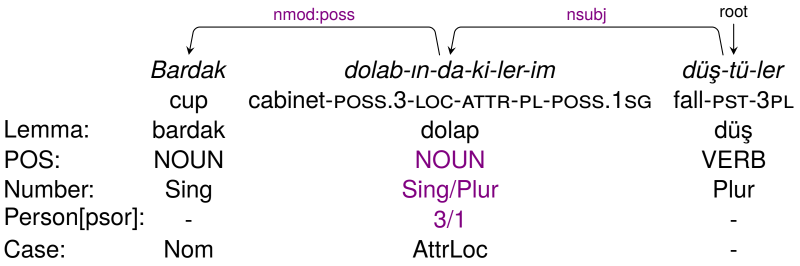
3.2. Layered features

An approach that would allow for annotation of different morphological features for the two referents of a pronominalised locative token is to use layered features.

While not currently used in this way in UD, layered features enable us to annotate more than one value on a feature key. Some Turkic treebanks have already employed layered features to annotate possessive marker on a nominal (cf. *dolab-ın-da-ki* and *bardak-lar-ım* in Figure 4, where psor in the brackets specifies that the Person key refers to the Person feature of the possessor). By extending their usage, it is possible to use layered features to specify which stem a feature key is referring to. The application of this approach on the example sentence is shown in Figure 7.

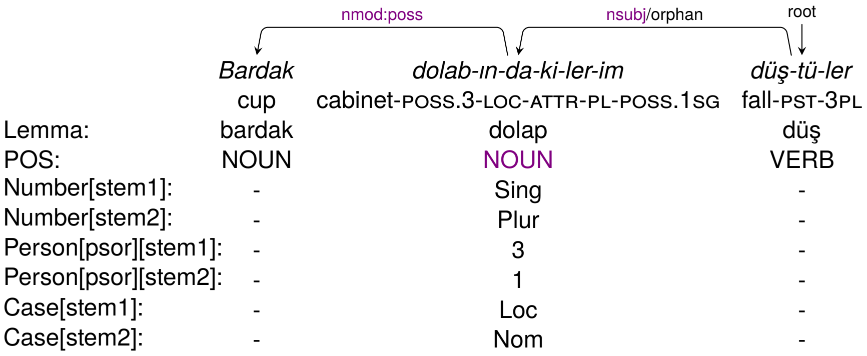
Advantages of this approach are that (i) we can annotate multiple features sharing the same key without splitting the word, (ii) layers can be recursively applied, (iii) layered features can be applied to languages without a derivational morpheme like *-ki* (e.g., some Tungusic, Quechuan, and Dargin languages), and (iv) it is compatible with the hierarchical annotation of morphology in UniMorph 4.0 (Batsuren et al., 2022).

This approach, however, fails to solve the dependency relation issues presented by having a single token: it is not clear which subword token is



	Bardak	*dolab-ın-da-ki-ler-im*	*düş-tü-ler*
	cup	cabinet-POSS.3-LOC-ATTR-PL-POSS.1SG	fall-PST-3PL
Lemma:	bardak	dolap	düş
POS:	NOUN	NOUN	VERB
Number:	Sing	Sing/Plur	Plur
Person[psor]:	-	3/1	-
Case:	Nom	AttrLoc	-

Figure 6: Analyzing *-ki* with no segmentation, with an `AttrLoc` case feature.



	Bardak	*dolab-ın-da-ki-ler-im*	*düş-tü-ler*
	cup	cabinet-POSS.3-LOC-ATTR-PL-POSS.1SG	fall-PST-3PL
Lemma:	bardak	dolap	düş
POS:	NOUN	NOUN	VERB
Number[stem1]:	-	Sing	-
Number[stem2]:	-	Plur	-
Person[psor][stem1]:	-	3	-
Person[psor][stem2]:	-	1	-
Case[stem1]:	-	Loc	-
Case[stem2]:	-	Nom	-

Figure 7: Analyzing *-ki* with no segmentation, using (extended) layered features.

the 'head', and which is the actual referent of the external dependency relation. There is also still only one POS.

In summary, there is a strong indication that the pronominal formed by *-ki* contains multiple syntactic words.

3.3. Splitting before *-ki*

Segmentation of the pronominalised forms solves the problems with conflicting features and dependencies, as well as the non-informativeness of the `orphan` relation. We consider two different ways (or locations) for segmenting these forms. The first option (Figure 8), which is used in some of the current treebanks (e.g., Türk et al., 2019; Marşan et al., 2022), considers the *-ki* morpheme as part of the second token.

This approach allows retaining all linguistic information packed in the *-ki* bearing forms:

- There are two referents: The possessor of the cup cabinet (third person singular) and the possessor of the items in the cabinet (first person singular). Both are clearly annotated in morphological features and POS tags in two subword tokens.
- The relationship between the two subwords is established (`nmod`, second subword being the head), and the external relationships between the *-ki* bearing form and other element(s) in the sentence are clear (the second subword being an `nsubj` dependent).
- The first subword can be annotated as taking part in other syntactic phenomena, such as compounding, independently of the full token. In our example here, the compound *bardak dolabı* is independent of (although a part of) the pronominal that is formed with *-ki*. Splitting the *-ki* bearing form into subwords allows illustrating such constructions more clearly.

In addition to enabling annotation of all morphological features and dependency relations, splitting before *-ki* prevents ending up with null morphemes (discussed in detail in §3.4). There are two disadvantages to this approach. Firstly, the current UD guidelines are not very supportive of subword tokenization, so this approach diverges from the UD framework to some extent. Secondly, due to the additional complexity, this approach can introduce noise or learnability issues for less sophisticated systems like shallow parsers.

3.4. Splitting after *-ki*

An alternative segmentation approach segments pronominalised locatives after *-ki*, as shown in Figure 9.

When splitting before *-ki*, the *-ki* morpheme is considered part of the pronominal 'word' (i.e., the part of the token representing the second referent). This can be viewed as inconsistent with the attributive use of *-ki*, where—regardless of whether or not *-ki* is best treated as an independent token—it is clear that *-ki* is not the lemma to which the

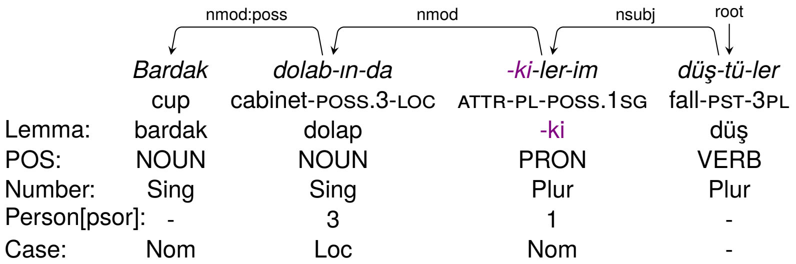


Figure 8: Possible analysis segmenting before *-ki*.

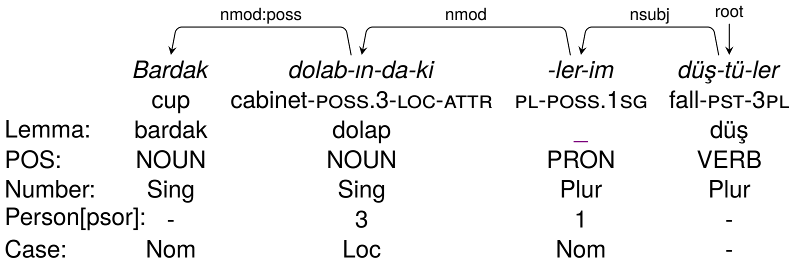


Figure 9: Possible analysis segmenting after *-ki*.

second set of morphological features belong. For example, in the annotation of the attributive use of *-ki* in Figure 4, the noun head of the *-ki* bearing form has the lemma *bardak*. However, in the annotation of an equivalent sentence with that noun absent and its morphology instead associated with the *-ki* bearing form, such as that in Figure 8, the pronoun head of the second referent (which could still be understood to refer to *bardak*), is now *-ki* according to the split-before approach. In other words, the *-ki* is associated with a different token in these two examples—and more broadly, in these two constructions: in an attributive construction, *-ki* is associated with the first participant, and in an equivalent pronominal construction, *-ki* is associated with the second participant.

The approach of splitting after *-ki*, then, is a way to avoid what might be seen as an inconsistency that arises when splitting before *-ki*. By segmenting pronominalised locatives immediately after *-ki*, the *-ki* morpheme remains with the first of the two tokens (the dependent and not the head) whether attributive or pronominal. This also unifies these two uses of *-ki* as a single phenomenon, with the addition of the phenomenon that allows the head noun to be absent in pronominal *-ki* forms.

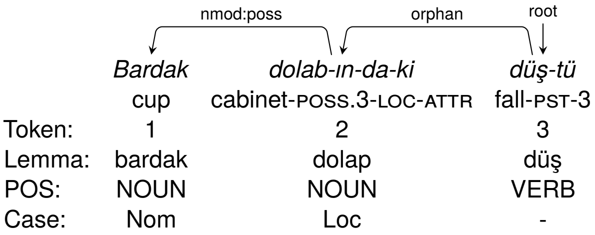
A major problem with this approach is that it requires an empty lemma, as well as an empty form when there are no additional affixes after *-ki*. Empty lemmas and forms are not allowed according to UD v2 annotation guidelines. While it would be possible not to annotate a second token (the pronoun / second referent) if it were empty, that would reduce the consistency of this approach, and still leaves the issue of having an empty lemma. Furthermore, as with segmenting before *-ki*, there may be limitations for less sophisticated automated annotation systems, although it is possible that systems capable of segmenting words into subword units would be able to handle one approach more easily than the other—an area for future investigation. Lastly, treating attributive and pronominal locatives uniformly may go against a generative syntax analysis of these two uses, where the attributive locative form is an ordinary member of the phrase (DP) containing the head noun, whereas the pronominal locative is cast directly into a DP with the accompanying morphology and has fewer layers between the two phrases.

3.5. Splitting after *-ki* with fallback

One problem with splitting after *-ki* is that null nodes would result in situations where there is no inflection, as in the sentence *Bardak dolabındaki düştü* 'The one on the cup cabinet fell'. This problem could be avoided with a fallback in such cases.

One option is to fall back to an `orphan` analysis, per Figure 10, signalling to downstream tasks that information is missing (specifically an elided [pronominal] element). Using the `orphan` relation has the disadvantages discussed in §3.1: it is not informative, and does not allow for annotation of multiple relations (although implies them) or multiple sets of features. However, examples of pronominalised locatives are not very frequent in existing corpora, and examples of pronominalised locatives with no further inflection are quite rare, so this approach would not result in excessive use of the `orphan` relation.

To include the elided information, enhanced dependencies may be used, as in Figure 11. En-



	Bardak	*dolab-ın-da-ki*	*düş-tü*
	cup	cabinet-POSS.3-LOC-ATTR	fall-PST-3
Token:	1	2	3
Lemma:	bardak	dolap	düş
POS:	NOUN	NOUN	VERB
Case:	Nom	Loc	-

Figure 10: Possible analysis segmenting after *-ki* with no morphology, with `orphan` fallback.

hanced dependencies are explicitly designed to present null nodes in cases of elision.[12] Use of enhanced dependencies has some drawbacks. If annotated even for just one example, the entire corpus needs to have enhanced dependencies annotated. Furthermore, most parsers, querying tools, and other applications of UD lack support for enhanced dependencies and ignore them. However, this approach does preserve the information lost in the accompanying standard dependency analysis.

4. Summary of approaches

The approaches described in Section 3, and their advantages and disadvantages are summarised in Table 1.

The first approach discussed, no-segmentation (§3.1), has the benefit of ease of tokenization. Even though state-of-the-art parsers may be successful in segmenting words into subword units, not having to split words has a clear advantage, especially in low-resource scenarios.[14] It also avoids empty word forms and empty lemmas that some of the approaches postulate. However, it fails to represent multiple sets of morphological features, and it does not allow a correct interpretation of the dependency relations the word participates in. Specifically, annotating in this way results in a situation where it is unclear which of the token's referents is the modifier of another head. Possible ways to remove the ambiguity would be to use the `orphan` relation (second row of Table 1) or an `AttrLoc` value for the case feature (third row of Table 1), both of which allow for differentiation of pronominalised locatives from other dependents with a similar relation to the head. However, `orphan` does not include any information regarding the syntactic function of the word in the sentence. With or without the `orphan` relation or an `AttrLoc` case feature, the no-segmentation approach does not resolve the issue of multiple, potentially conflicting sets of morphological features assigned to a single syntactic word.

A possible solution (described in §3.2) that allows expressing multiple sets of morphological features is to make use of layered features as exemplified in Figure 7. Although this uses the UD layered features in an unorthodox way,[15] it enables specification of multiple sets of morphological features, and, with the use of the `orphan` relation, pronominalised locatives can also be differentiated from other dependents with a similar relation to their head. However, as noted earlier, it does not allow identifying the dependency relations correctly. It still leaves it unclear which part of the word is modified by a modifier, and which part is a modifier to another head. Another downside is, perhaps, the complexity: such feature sets and relations are likely to be difficult to learn for parsers, and the treebank queries for relevant features/structures are likely to be misled or miss the relevant items due to the idiosyncratic nature of the annotations.

Both segmentation options resolve the main concerns with the pronominal construction: the appropriate features are easily assigned to each syntactic word, and the dependents can modify the correct syntactic word without ambiguity. The relation between the pronominal and its head is also clearer. The disadvantage of splitting before *-ki* (§3.3) is the inconsistency with the attributive use. This approach suggests either splitting *-ki* in attributive usage without any clear motivation—in which case it is still not the lemma of the modified noun's morphological features as in the pronominal treatment—or treating attributive and pronominal cases differently.[16] The disadvantage of splitting after *-ki* (§3.4) is the introduction of empty lemmas, and empty forms when no further affixes are attached after *-ki*. Since empty forms are not allowed in the current basic UD dependencies, this approach would require a substantial modification to the UD guidelines. Splitting after *-ki* with fallback (§3.5) solves the issue of empty lemmas but requires the use of enhanced dependencies

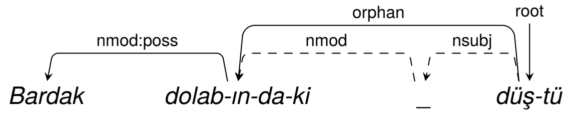
[12] Per https://universaldependencies.org/v2/enhanced.html.

[13] Empty forms and lemmas would only occur in enhanced dependencies annotation, where they are permissible.

[14] We intend to investigate this empirical question in future research.

[15] E.g., introducing multi-dimensional layers, and layers indexed by ordinals.

[16] Which may also result in difficulties with the automated processing.



	Bardak	*dolab-ın-da-ki*	_	*düş-tü*
	cup	cabinet-POSS.3-LOC-ATTR		fall-PST-3
Token:	1	2	2.1	3
Lemma:	bardak	dolap	_	düş
POS:	NOUN	NOUN	PRON	VERB
Case:	Nom	Loc	Nom	-

Figure 11: Possible analysis segmenting after *-ki* with no morphology, with enhanced dependencies fallback.

Approach	No empty forms	No empty lemmas	2 sets of features	Deprels for 2 referents	Consistent with attributive use	Easy querying	No need for subword-aware parser
No-segmentation	✔	✔	✘	✘	✘	✔	✔
`orphan` relation	✔	✔	✘	✘	✘	✔	✔
`AttrLoc` feature	✔	✔	✘	✘	✘	✔	✔
Layered features	✔	✔	✔	✘	✘	✘	✔
Splitting before *-ki*	✔	✔	✔	✔	✘	✔	✘
Splitting after *-ki*	✘/✔	✘	✔	✔	✔	✔	✘
Splitting after, enhanced dependencies fallback	(✔)[13]	(✔)[13]	✔	✔	✔	✘	✘

Table 1: A summary of the advantages and disadvantages in the discussed approaches.

framework, which introduces a new set of challenges including compatibility issues for existing UD tools.

5. Concluding remarks

The authors currently consider splitting pronominalised locatives before *-ki* a best compromise, and recommend this for annotation of Turkic treebanks, although with a caveat.

While the authors agree with one another that segmentation is needed to properly capture these constructions, opinions differ as to which approach is ideal. Proponents of splitting the pronominalised locative before *-ki* do not believe that it is a problem for the approach to be inconsistent with the treatment of the attributive locative due to a generative syntax view that they are in fact distinct. Proponents of splitting the pronominalised locative after *-ki* realise that it would take a major change to current UD guidelines for this approach to be viable, and while finding splitting before *-ki* somewhat unsatisfactory, accept that it may be the current best compromise.

The issue of pronominalised locatives is just one of many specific issues where consistent UD annotation guidelines are needed for Turkic languages. This issue is also relevant to the UD (and UniDive) community at large. By bringing awareness to this issue and discussing it in depth, we hope that new annotation projects for languages with similar phenomena will be eased, and that our efforts will lead to improved overall quality of corpora and annotation guidelines.

6. Acknowledgements

We are grateful to Gülnara Karasawa for her help with creating the Tatar dataset for this study.

We thank UniDive, the COST Action CA21167, and Istanbul Technical University for supporting the organization of the workshop in September 2023 that made the present work and future collaboration towards a unified annotation of Turkic UD treebanks possible. We also thank all participants of the workshop for fruitful discussion and suggestions.

7. Bibliographical References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam

Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Özlem Çetinoğlu and Çağrı Çöltekin. 2019. Challenges of annotating a code-switching treebank. In *Proceedings of the 18th international workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 82–90.

Özlem Çetinoğlu and Çağrı Çöltekin. 2022. Two languages, one treebank: building a Turkish–German code-switching treebank and its challenges. *Language Resources and Evaluation*, pages 1–35.

Çağrı Çöltekin. 2015. A grammar-book treebank of Turkish. In *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 35–49.

Çağrı Çöltekin, A Doğruöz, and Özlem Çetinoğlu. 2022. Resources for Turkish natural language processing: A critical survey. *Language Resources and Evaluation*.

Mehmet Oguz Derin and Takahiro Harada. 2021. Universal Dependencies for Old Turkish. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 129–141, Sofia, Bulgaria. Association for Computational Linguistics.

Marhaba Eli, Weinila Mushajiang, Tuergen Yibulayin, Kahaerjiang Abiderexiti, and Yan Liu. 2016. Universal dependencies for Uyghur. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 44–50, Osaka, Japan. The COLING 2016 Organizing Committee.

Federica Gamba and Daniel Zeman. 2023a. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Federica Gamba and Daniel Zeman. 2023b. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. London: Routledge.

Martin Haspelmath and Andrea D. Sims. 2010. *Understanding Morphology*, second edition. Understanding Language. Taylor & Francis.

Aida Kasieva, Gulnura Dzhumalieva, Anna Thompson, Murat Jumashev, Bermet Chontaeva, and Jonathan Washington. 2023. Issues of Kyrgyz syntactic annotation within the Universal Dependencies framework. In *Proceedings of the XI International Conference on Computer Processing of Turkic Languages (TurkLang 2023)*.

Bonnie Krejci and Lelia Glass. 2015. The Kazakh noun/adjective distinction. In *Proc. of the 9th Workshop on Altaic Formal Linguistics (WAFL9)*, pages 47–58, Cambridge, MA. MITWPL.

Aibek Makazhanov, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev. 2015a. Syntactic annotation of Kazakh: Following the universal dependencies guidelines. a report. In *Proceedings of the 3rd International Conference on Computer Processing in Turkic Languages (TurkLang 2015)*, pages 338–350.

Aibek Makazhanov, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev. 2015b. Syntactic annotation of Kazakh:

Following the universal dependencies guidelines. a report. In *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pages 338–350.

Büşra Marşan, Salih Furkan Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Betül Özateş, Suzan Üsküdarlı, Arzucan Özgür, Tunga Güngör, and Balkız Öztürk. 2022. Enhancements to the boun treebank reflecting the agglutinative nature of turkish. In *The Proceedings of the ALTNLP2022 The International Conference and workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing*, pages 71–80.

Tatiana Merzhevich and Fabrício Ferraz Gerardi. 2022. Introducing YakuToolkit. Yakut treebank and morphological analyzer. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 185–188, Marseille, France. European Language Resources Association.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan. The COLING 2016 Organizing Committee.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2019. Improving the annotations in the Turkish Universal Dependency treebank. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 108–115, Paris, France. Association for Computational Linguistics.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2022. Resources for turkish dependency parsing: Introducing the boun treebank and the boat annotation tool. *Language Resources and Evaluation*, pages 1–49.

Francis Tyers and Jonathan Washington. 2015. Towards a free/open-source universal-dependency treebank for Kazakh. In *Proceedings of the 3rd International Conference on Computer Processing in Turkic Languages (TurkLang 2015)*, pages 276–289.

Francis Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An assessment of Universal Dependency annotation guidelines for Turkic languages. In *5th International Conference on Turkic Language Processing (TURKLANG 2017)*, pages 356–377.

Jonathan N. Washington, Francis M. Tyers, and Ilnar Salimzianov. 2022. Non-finite verb forms in Turkic exhibit syncretism, not multifunctionality. *Folia Linguistica*, 56(3):693–742.

Jonathan North Washington and Francis Morton Tyers. 2019. Delineating Turkic non-finite verb forms by syntactic function. In *Proceedings of the Workshop on Turkic and Languages in Contact with Turkic*, volume 4, pages 115–129.

Amir Zeldes and Nathan Schneider. 2023. Are UD treebanks getting more consistent? a report card for English UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 58–64, Washington, D.C. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Çağrı Çöltekin. 2016. (When) do we need inflectional groups? In *Proceedings of The First International Conference on Turkic Computational Linguistics*.

8. Language Resource References

Ibrahim Benli. 2023. *Kyrgyz KTMU Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Cesur, Neslihan and Kuzgun, Aslı and Yıldız, Olcay Taner and Marşan, Büşra and Kara, Neslihan and Arıcan, Bilge Nas and Özçelik, Merve and Aslan, Deniz Baran. 2023a. *Turkish Penn Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Cesur, Neslihan and Kuzgun, Aslı and Yıldız, Olcay Taner and Marşan, Büşra and Kuyrukçu, Oğuzhan and Arıcan, Bilge Nas and Sanıyar, Ezgi and Kara, Neslihan and Özçelik, Merve. 2023b. *Turkish FrameNet Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Eli, Marhaba and Zeman, Daniel and Tyers, Francis. 2023. *Uyghur UDT Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kuzgun, Aslı and Cesur, Neslihan and Yıldız, Olcay Taner and Kuyrukçu, Oğuzhan and Marşan, Büşra and Arıcan, Bilge Nas and Kara, Neslihan and Aslan, Deniz Baran and Sanıyar, Ezgi and Asmazoğlu, Cengiz. 2023a. *Turkish Tourism Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kuzgun, Aslı and Cesur, Neslihan and Yıldız, Olcay Taner and Kuyrukçu, Oğuzhan and Yenice, Arife Betül and Arıcan, Bilge Nas and Sanıyar, Ezgi. 2023b. *Turkish Kenet Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Köse, Mehmet and Yıldız, Olcay Taner. 2023. *Turkish Atis Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Makazhanov, Aibek and Washington, Jonathan North and Tyers, Francis. 2023. *Kazakh KTB Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Marşan, Büşra and Akkurt, Salih Furkan and Türk, Utku and Atmaca, Furkan and Özateş, Şaziye Betül and Berk, Gözde and Bedir, Seyyit Talha and Köksal, Abdullatif and Başaran, Balkız Öztürk and Güngör, Tunga and Özgür, Arzucan. 2023. *Turkish BOUN Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Merzhevich, Tatiana and Gerardi, Fabrício Ferraz. 2023. *Yakut YKTDT Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Taguchi, Chihiro. 2023. *Tatar NMCTT Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Türk, Utku and Özateş, Şaziye Betül and Marşan, Büşra and Akkurt, Salih Furkan and Çöltekin, Çağrı and Cebiroğlu Eryiğit, Gülşen and Gökırmak, Memduh and Kaşıkara, Hüner and Sulubacak, Umut and Tyers, Francis. 2023. *Turkish IMST Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Uszkoreit, Hans and Macketanz, Vivien and Burchardt, Aljoscha and Harris, Kim and Marheinecke, Katrin and Petrov, Slav and Kayadelen, Tolga and Attia, Mohammed and Elkahky, Ali and Yu, Zhuoran and Pitler, Emily and Lertpradit, Saran and Cetin, Savas and Popel, Martin and Zeman, Daniel and Tyers, Francis and Çöltekin, Çağrı and Türk, Utku and Atmaca, Furkan and Özateş, Şaziye Betül and Köksal, Abdullatif and Başaran, Balkız Öztürk and Güngör, Tunga and Özgür, Arzucan. 2023. *Turkish PUD Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Çetinoğlu, Özlem and Çöltekin, Çağrı. 2023. *Turkish German SAGT Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Çöltekin, Çağrı. 2023. *Turkish GB Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A. UD Turkic Treebanks

There are currently UD treebanks for Kazakh, Kyrgyz, Tatar, Turkish, Uyghur, Yakut, and Old Turkish, and a treebank annotating sentences with Turkish-German code switching. All languages except Turkish are represented with a single treebank, while Turkish has 9 treebanks. Table 2 lists the treebanks currently released in the UD repositories as of UD version 2.13.

	sent	tok	multi	types	ltypes	pos	rel	feat
Kazakh/KTB (Tyers and Washington, 2015; Makazhanov et al., 2015a) (Makazhanov et al., 2023)	1078	10536	41	4642	2433	17	36	9
Kyrgyz/KTMU (Benli, 2023)	781	7451	0	3474	2305	13	26	8
Old Turkish/Tonqq (Derin and Harada, 2021)	20	158	0	75	2	13	19	0
Tatar/NMCTT (Taguchi, 2023)	148	2280	0	1264	843	14	28	7
Turkish/Atis (Köse and Yıldız, 2023)	5432	45907	0	2133	995	13	36	7
Turkish/BOUN (Türk et al., 2022; Marşan et al., 2022) (Marşan et al., 2023)	9761	125212	3374	37052	12649	16	46	7
Turkish/FrameNet (Cesur et al., 2023b)	2698	19223	0	8403	3905	15	30	7
Turkish/GB (Çöltekin, 2015) (Çöltekin, 2023)	2880	17177	371	5517	2074	16	42	7
Turkish/IMST (Sulubacak et al., 2016) (Türk et al., 2023)	5635	58096	1639	18541	5960	14	40	10
Turkish/Kenet (Kuzgun et al., 2023b)	18687	178658	0	49156	15343	15	34	7
Turkish/Penn (Cesur et al., 2023a)	16396	183555	0	37765	14977	15	36	9
Turkish/PUD (Zeman et al., 2017) (Uszkoreit et al., 2023)	1000	16881	346	7646	4598	16	38	4
Turkish/Tourism (Kuzgun et al., 2023a)	19830	91152	0	4961	2170	15	33	13
Turkish-German/SAGT (Çetinoğlu and Çöltekin, 2022) (Çetinoğlu and Çöltekin, 2023)	2184	37227	290	7094	3836	17	45	12
Uyghur/UDT (Eli et al., 2016) (Eli et al., 2023)	3456	40236	0	12067	2908	16	45	15
Yakut/YKTDT (Merzhevich and Ferraz Gerardi, 2022) (Merzhevich and Gerardi, 2023)	299	1460	1	688	405	14	26	6

Table 2: Basic statistics on current UD treebanks (as of UD version 2.13). *sent*: number of sentences, *tok*: number of tokens, *multi*: number of multi-word tokens, *types*: number of word types, *ltypes*: number of lemma types, *pos*: number of POS tags used, *rel*: number of dependency relations used (including language/treebank specific relations), *feat*: number of morphological features used.

Besides existing treebanks, the UD web page also reports Uzbek, Ottoman Turkish and yet another Turkish treebank in preparation. We are also aware of new treebanks in preparation for Kyrgyz (Kasieva et al., 2023), Azerbaijani and Kumyk.

BERT-based Idiom Identification using Language Translation and Word Cohesion

Arnav Yayavaram*, Siddharth Yayavaram*, Prajna Upadhyay, Apurba Das
BITS Pilani Hyderabad, India
{f20213117, f20213116, prajna.u, apurba}@hyderabad.bits-pilani.ac.in

Abstract

An idiom refers to a special type of multi-word expression whose meaning is figurative and cannot be deduced from the literal interpretation of its components. Idioms are prevalent in almost all languages and text genres, necessitating explicit handling by comprehensive NLP systems. Such phrases are referred to as Potentially Idiomatic Expressions (PIEs) and automatically identifying them in text is a challenging task. In this paper, we propose using a BERT-based model fine-tuned with custom objectives, to improve the accuracy of detecting idioms in text. Our custom loss functions capture two important properties (word cohesion and language translation) to distinguish PIEs from non-PIEs. We conducted several experiments on 7 datasets and showed that incorporating custom objectives while training the model leads to substantial gains. Our models trained using this approach also have better sequence accuracy over DISC, a state-of-the-art PIE detection technique, along with good transfer capabilities. Our code and datasets can be downloaded from https://github.com/siddharthyayavaram/BERT-Based-Idiom-Detection

Keywords: idioms, multi-word expressions, word cohesion, language translation, loss function

1. Introduction

An idiom refers to a special type of multi-word expression (Baldwin and Kim, 2010) whose meaning is figurative and cannot be deduced from the literal interpretation of its components. Idioms often exhibit peculiar behavior by violating selection restrictions or altering the default semantic roles of syntactic categories. Consequently, they pose significant challenges for Natural Language Processing (NLP) systems. Idioms are prevalent in almost all languages and text genres, necessitating explicit handling by comprehensive NLP systems. We refer to these phrases as *potentially idiomatic expressions (PIEs)* to account for the contextual semantic ambiguity in their expression. Better detection of PIEs can enhance numerous machine translation tasks.

Techniques to automatically detect and identify PIEs need to do many tasks accurately – $i)$ automatically detect if an idiomatic expression is present in a sentence (Briskilal and Subalalitha, 2022; Tan and Jiang, 2021; Liu and Hwa, 2019), $ii)$ if yes, identify the idiomatic tokens (Zeng and Bhat, 2021, 2022). Both of these are challenging tasks. For instance, in the sentence“Oh — for about four years, on and off, he said vaguely”, the potentially idiomatic expression “on and off” is used figuratively, whereas, it is used literally in the sentence “Participate in training, both on and off station”. Existing techniques for idiom detection rely on syntactic patterns, knowing the PIE being classified correctly, and lack generalization. In this paper, we address the above-mentioned problems and show that improvement in $i)$ improves $ii)$ substantially.

We employ a BERT-based fine-tuning approach with custom objectives to improve accuracy on all 3 tasks. We define our objectives in Section 4.2 based on language translation and word cohesion.

Our salient contributions are:
1: Introduction of a language translation-based metric to detect the presence of idioms.
2: A novel loss function to selectively penalize examples using sentence translation and word cohesion that can be used with any architecture for idiom detection.
3: Our models trained with custom loss functions exhibit improved generalization capabilities, evident in identifying unseen PIEs.

2. Related Work

MWE, short for Multi Word Expressions are notable collocations with multiple words, for instance “all at once” or “look something up”. (Baldwin and Kim, 2010; Constant et al., 2017). IEs (Idiomatic Expressions), are a subset of MWEs, which exhibit non-compositionality (Baldwin and Kim, 2010; Fadaee et al., 2018; Liu et al., 2017; Biddle et al., 2020). Metaphors, such as “heart of gold” and “night owl” compare unrelated things implicitly. While some MWEs and IEs use metaphorical figuration, not all metaphors are IEs; they can be direct comparisons with single words (e.g., “I am titanium”). In this paper, we study IEs.

*Equal Contribution

IE Classification broadly falls under two categories – standalone phrase classification and context-based classification. Standalone classification tasks decide if a phrase could be used as an idiom without specifically considering its context (Fazly and Stevenson, 2006; Shutova et al., 2010; Tabossi et al., 2008, 2009; Reddy et al., 2011; Cordeiro et al., 2016) as opposed to context-based idiom classification techniques which take into account the entire sentence to detect the presence of idiom (Peng et al., 2014; Nedumpozhimana et al., 2022; Peng and Feldman, 2017; Tan and Jiang, 2021; Verma and Vuppuluri, 2015; Briskilal and Subalalitha, 2022; Liu and Hwa, 2019). Earliest known context-based phrase classification techniques developed per idiom classifiers, which are not scalable (Liu and Hwa, 2017). Context-based phrase classification techniques can additionally detect which tokens are idiomatic/nonidiomatic (Zeng and Bhat, 2021; Salton et al., 2016; Zeng and Bhat, 2022). Typically, the latter is dependent on the former task – only if an idiom is detected to be present in a sentence, does the classification of idiomatic and non-idiomatic tokens follow. Efforts to build complementary resources to support this task include constructing a knowledge graph (Zeng et al., 2023) and an information retrieval system to search for idiomatic expressions (Hughes et al., 2021).

Detecting idioms in the text has also become popular in non-English languages. In (Itkonen et al., 2022), authors leverage various models provided by HuggingFace in conjunction with the standard BERT model for the idiom detection task in English, Portuguese, and Galician. They emphasize on feature engineering using traits that define idiomatic expressions. These additional features result in enhancements compared to the baseline performance. In (Tedeschi et al., 2022), a multilingual transformer based model and a dataset of idioms in 10 languages is presented. A rule-based intra-sentential idiom detection system in Hindi was presented in (Priyanka and Sinha, 2014).

3. Problem Statement

We are given the following:

- A sentence $\mathcal{S}$ with n tokens $w_1, w_2, \ldots, w_n$, where each w_i represents a tokenized unit. $\mathcal{S}$ is an syntactic ordering over w_i's.
- Labels $\mathcal{L} = \{\texttt{I}, \texttt{NI}\}$ where `I` and `NI` represent `<idiom>` and `<not idiom>` (or literal) classes, respectively.

This labelling produces a sequence of class labels $Z = z_1, z_2, \ldots, z_n$ where $z_i = f(w_i)$. The high-level objective of this work is to learn the function $f(\cdot)$

- A successful prediction occurs when an idiomatic subsequence $w_{i:j}$ is identified in S, and the corresponding labels $z_{i:j}$ are labelled as `I`. There can be more than one such subsequences.
- If the subsequence $w_{i:j}$ is literal, all corresponding labels $z_{i:j}$ are `NI`.
- If the sentence lacks an idiom, all $z_{1:n}$ are categorized as `NI`.

4. Methodology

4.1. BERT-based Idiom Identification

Figure 1 shows the high-level architecture of our method. Our loss functions are implemented over BERT (Devlin et al., 2018), a pre-trained transformer-based model developed by Google. Due to its effectiveness in capturing context and semantics for various NLP tasks, we re-use its pretrained architecture for fine-tuning our model using binary cross-entropy loss. Despite its success, cross entropy loss is sensitive to outliers and class-imbalance. We observe class imbalance in idiom classification where the label `I` is far less frequent than label `NI` leading to poor accuracy for `I` tokens. To fix this, we propose to use language translation and word cohesion to manipulate the loss. In the following sections, we define two novel loss functions for the task of idiom token classification. The merit of our work lies in the fact that these custom loss functions can be used with **any** architecture.

4.2. Language Translation and Cohesion for Idioms

4.2.1. Translation-based Loss Function

An important property exhibited by an idiom is the difference between its literal and actual meaning. However, a phrase that is an idiom in language L_1 is improbable to be an idiomatic phrase in another language L_2.For example, take the English idiom, “raining cats and dogs”, its Hindi translation is “भारी वर्षा”, which when translated back to English gives “heavy rain” which is the meaning of our initial idiom but is quite different lexically. Let $\mathcal{S}_{L_1}$ denote a sentence containing an idiom in language L_1, $\mathcal{S}_{L_1 \to L_2}$ a translation of $\mathcal{S}_{L_1}$ in L_2, and $\mathcal{S}_{L_1 \rightleftarrows L_2}$ a translation of $\mathcal{S}_{L_1 \to L_2}$ back to L_1. When $\mathcal{S}_{L_1}$ is translated to $\mathcal{S}_{L_1 \to L_2}$, the idiomatic tokens in $\mathcal{S}_{L_1}$ will be expressed through their actual meaning in $\mathcal{S}_{L_1 \to L_2}$ because of a lack of corresponding idiom in L_2. Re-translating it to L_1 will force the idiom to be expressed with its actual meaning in $\mathcal{S}_{L_1 \rightleftarrows L_2}$. Lexically, the actual meaning of an idiom and the surface form of an idiom differ substantially from each other. We employ this simple trick to detect

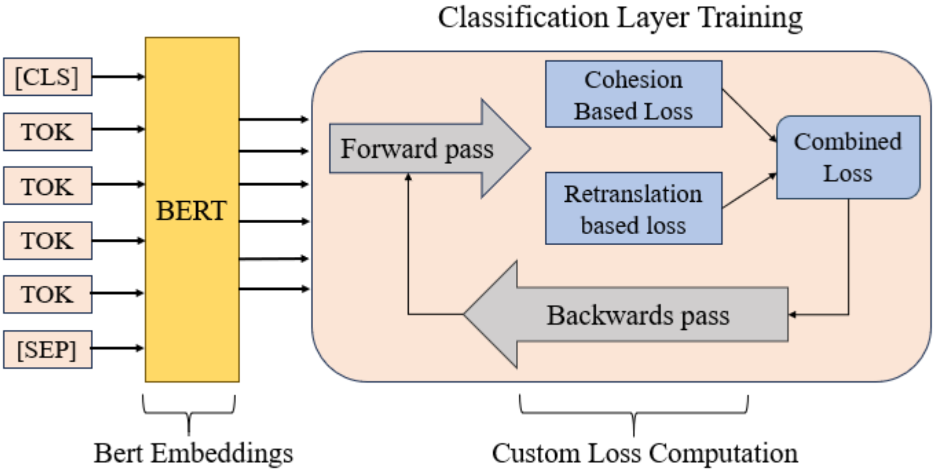


Figure 1: Architecture of our proposed method

the presence of an idiom in a sentence – if $\mathcal{S}_{L_1 \rightleftarrows L_2}$ and $\mathcal{S}_{L_1}$ differ lexically by some margin, $\mathcal{S}$ is likely to contain an idiom. A sentence that contains no idiom is likely to have the same lexical representation in the original and back-translated sentence. We leverage the METEOR (Banerjee and Lavie, 2005) metric to quantify this observation by computing a score to reflect the lexical and syntactic similarity between the translated and reference sentences. METEOR incorporates a penalty mechanism for longer matches by organizing system translation unigrams mapped to reference translation unigrams into minimal chunks. These chunks consist of adjacent unigrams in the system translation that align with adjacent unigrams in the reference translation. Longer n-grams result in fewer chunks. In the extreme case of a complete match, only one chunk exists, while in the absence of bigram or longer matches, the number of chunks equals the count of unigram matches. An alignment is created between the system translation and the reference translation by mapping unigrams based on different criteria, such as exact match, stemming, or synonymy. The alignment is formed by selecting the most extensive subset of unigram mappings, ensuring that each unigram maps to at most one unigram in the other string. The chosen alignment is the one with the fewest "unigram mapping crosses", which occur when lines connecting mapped unigrams intersect in a vertical arrangement of the two strings.

$$\text{Unigram Precision:} \qquad \mathcal{P} = \frac{N_{\text{correct}}}{N_{\text{backtrans}}}$$

$$\text{Unigram Recall:} \qquad \mathcal{R} = \frac{N_{\text{correct}}}{N_{\text{original}}}$$

Here, N_{correct} represents the number of correctly mapped unigrams, $N_{\text{backtrans}}$ represents the total number of unigrams in the back-translated sentence, and N_{original} represents the total number of unigrams in the original sentence.

$$\text{Harmonic Mean:} \qquad \mathcal{F}_{\text{mean}} = \frac{10 \cdot \mathcal{P} \cdot \mathcal{R}}{\mathcal{R} + 9 \cdot \mathcal{P}}$$

$$\text{Penalty} = 0.5 \times \left(\frac{C}{U}\right)^3$$

where C represents the number of chunks and U represents the number of unigrams matched.

$$\text{Score} = \mathcal{F}_{\text{mean}} \times (1 - \text{Penalty})$$

It evaluates the quality of a translation by comparing it to one or more reference translations. METEOR considers various factors such as unigram precision, recall, and alignment errors to compute a score that reflects the lexical and syntactic similarity between the translated and reference sentences. For instance, the sentence "The early morning flight required them to hit the sack much earlier than usual", is translated into Italian "Il volo mattutino li obbligava a coricarsi molto prima del solito.", and its back-translation to English "The morning flight forced them to go to bed much earlier than usual.", the idiomatic usage causes a large syntactic change during back-translation which will lead to a high alignment error term and comparatively lower METEOR score of 0.5919.

During the training of the BERT-based model for idiom recognition, the translation-based loss function incorporates the METEOR score as a penalty term. If the METEOR score falls below a certain threshold, it indicates that the back-translation process has significantly altered the original sentence, which we posit is due to the presence of idiomatic expressions.

$$\mathcal{L}_{retranslation} = \mathcal{L}(1 + \lambda_1 \mathbb{1}(\mathcal{MS} < \lambda_2)) \qquad (1)$$

where $\mathcal{MS}$ is the meteor score for the sentence, $\mathcal{L}$ is the original binary cross entropy loss, and $\mathbb{1}(|\mathcal{MS}| < \lambda_2)$ is an indicator function. It takes a value of 1 if it is low (< λ_2) which scales the loss λ_1 times. Otherwise, it defaults to regular loss $\mathcal{L}$.

By increasing the loss for examples where idioms are not accurately retained through back-translation, the model is encouraged to better understand and retain the meaning of idiomatic expressions. This, in turn, leads to improved performance metrics such as precision and recall, as the model becomes more adept at recognizing and appropriately handling idiomatic language during inference, resulting in better generalization to unseen data.

4.2.2. Cohesion based Loss Function

Idioms exhibit a lack of semantic compositionality or *cohesion* among its words also reported in earlier work (Baldwin and Kim, 2010). Given a sentence $\mathcal{S}$ where all tokens in the subsequence $w_{i:j}$ are tagged as $\texttt{I}$, we quantify the cohesion $C_{\mathcal{S}}$ among the words in $\mathcal{S}$ using Equation 2. It captures the mean similarity among the words in $\mathcal{S}$.

$$C_{\mathcal{S}} = \frac{1}{N} \sum_{w_i, w_j \in \mathcal{S}, i \neq j} \text{sim}(V(w_i), V(w_j)) \quad (2)$$

where $V(w_i)$ is an embedding vector for w_i, N is the total number of pairs of tokens in $\mathcal{S}$, and sim$(V(w_i), V(w_j))$ captures semantic similarity between w_i and w_j using $V(w_i)$ and $V(w_j)$. The 'sim' score is computed as the cosine similarity between the high dimensional vectors for each word. Its values range from -1 to 1, where 1 indicates high similarity and lexical cohesion, 0 represents dissimilar or orthogonal tokens, and -1 suggests that the vectors are in opposite directions. Similarly, we compute $C_{\mathcal{S}'}$, where $\mathcal{S}'$ is a sentence with the idiom tokens $w_{i:j}$ removed. The key idea is if $C_{\mathcal{S}'}$ is substantially higher than $C_{\mathcal{S}}$, then the $\mathcal{S}$ is highly likely to contain an idiomatic phrase. This follows from the intuition that idiomatic tokens are remotely related semantically to non-idiomatic tokens in $\mathcal{S}$ and their removal should increase the cohesion score.

We introduce this idea as loss during the fine-tuning objective. By penalizing examples with $\texttt{I}$ classifications that are not likely to contain idioms, it is guiding the model to differentiate between idiomatic and non-idiomatic sentences. Our cohesion-based loss function $\mathcal{L}_{cohesion}$ is expressed in Equation 3.

$$\mathcal{L}_{cohesion} = \mathcal{L}(1 + \lambda_3 \mathbb{1}(|\mathcal{C}_{S_1} - \mathcal{C}_{S_2}| > \lambda_4)) \quad (3)$$

where $\mathcal{C}_{S_1}$ and $\mathcal{C}_{S_2}$ are the cohesion scores for sentence $\mathcal{S}$ without and with the target idiom, respectively, $\mathcal{L}$ is the original binary cross entropy loss, and $\mathbb{1}(|\mathcal{C}_{S_1} - \mathcal{C}_{S_2}| > \lambda_4)$ is an indicator function. It takes a value of 1 if there is sufficient difference between cohesion scores $\mathcal{C}_{S_1}$ and $\mathcal{C}_{S_2}$ (> λ_4) which scales the loss λ_3 times. Otherwise, it defaults to regular loss $\mathcal{L}$.

4.3. Final Loss

The final loss is a linear combination of $\mathcal{L}_{retranslation}$ and $\mathcal{L}_{cohesion}$.

$$\mathcal{L}_{final} = \tau_1 \mathcal{L}_{retranslation} + \tau_2 \mathcal{L}_{cohesion} \quad (4)$$

τ_1 and τ_2 $(0 \leq \tau_i \leq 1)$ are parameters to control the effect of both losses. These parameters depend on the accuracy of $C_{\mathcal{S}}$ and $\mathcal{MS}$, which is determined by the quality of underlying embedding vectors (Equation 2) and translation API used. More weight can be given to the more accurate value.

5. Experiments

In this section, we present an empirical evaluation of our models on synthetic and real-world datasets to show the capabilities of our custom loss functions. We also compare our models with state-of-the-art techniques like DISC (Zeng and Bhat, 2021) — and we observed that using our custom loss functions leads to improved accuracies.

5.1. Experimental Setup

For training and testing our models, we make use of a 32×2 cores AMD EPYC5037532 server with 1 TB of RAM, and 8x A100 SXM4 80GB504. We used `bert-based-uncased` as our base model which we finetune.

In our experiments, we adapted the pre-trained `bert-base-uncased` (Devlin et al., 2018) model from Hugging Face [1] and proceed with fine-tuning. We selected this model primarily for its moderate size, which strikes a balance between performance and computational efficiency. Additionally, the "uncased" variant simplifies text processing by disregarding case sensitivity, making it faster to process. These factors make it a practical choice for token classification tasks without compromising performance. We selected Hindi as the language we translate to.

We partitioned each dataset into training (80%), validation (10%), and test sets (10%). Next, we applied a BERT tokenizer on the texts for generating tokens. This step is essential because it transforms the raw text data for input into the BERT model, which operates at the token level rather

[1] https://huggingface.co/docs/trl/en/models

than the word level. By converting words into token IDs, the tokenizer enables the model to understand and process the text effectively.

After tokenization, we aligned the labels with tokens to establish the correspondence between input tokens and the corresponding class labels. This alignment ensures that each token in the input text is associated with the correct entity label, allowing the model to learn the mapping between tokens and entity types during training. The alignment function handles cases where words are split into subwords by the tokenizer, ensuring that the labels are assigned appropriately to each token, even in the presence of subwords. We excluded special tokens representing separation between sentences and the start of the sentence from the training loss calculation by assigning them special labels.

We trained our model for three epochs, observing a sharp drop in loss over each epoch with a learning rate of ‘2e-5‘. The training and evaluation batch sizes were set to 16. Weight decay was set to ‘0.01‘ to avoid overfitting. We set λ_1, λ_2, λ_3, and λ_4 all to 999, and τ_1 and τ_2 to 0.01. We repeated our experiments for three seeds and reported average accuracy values (Table 2). Additionally, it is worth noting that we observe minimal deviation in accuracy across different random seeds which underscores the robustness of the results.

5.2. Baselines

`BERT-based approach (without custom loss).` We fine-tuned the BERT model with binary cross-entropy loss.

`BERT-based approach (with loss).` We used translation, cohesion, and combination losses (described in Section 4.2) to fine-tune our BERT model.

`DISC.` The DISC model is based on BERT, it uses contextualized and static embeddings to encode tokens using attention, and performs token-level literal/idiomatic classification, resulting in the final output. We compare DISC with our models on the Sequence Accuracy metric described in Section 5.4.

5.3. Datasets

Table 1 describes statistics of all the datasets we have used.

`1) magpie.` Derived from the British National Corpus (BNC) and annotated for idiomatic expressions (PIEs)(Haagsma et al., 2020)(Consortium, 2007), the MAGPIE corpus comprises 1756 PIEs across various syntactic patterns, alongside 56622 annotated instances (32.24 per PIE). We focused on fully figurative or literal samples, ensuring unambiguous tagging reflected in confidence scores. The resulting dataset includes approximately 37000 complete sentences, excluding those longer than 50 tokens.

Dataset	total number of sentences	#idioms	#sentences containing idioms	average sentences per idiom
`MAGPIE`	36192	1727	27727	16.05
`VNC-Tokens`	2571	48	2111	43.97
`theidioms`	7380	1606	7830	4.87
`formal`	3136	358	3136	8.76
`gtrans`	440	22	440	20
`gpt+gtrans`	880	22	440	20
`theidioms 1-1`	1606	1606	1606	1

Table 1: Statistics of the datasets used

`2) VNC-Tokens Dataset.` The VNC (Verb-Noun Combinations) corpus, sourced from the British National Corpus (BNC)(Cook et al., 2008)(Consortium, 2007), comprises 53 potentially idiomatic expressions (PIEs) with about 2500 annotated sentences, categorized as literal or figurative. Using regular expression libraries and the `NLTK` library [2], we annotated tokens as idiomatic or non-idiomatic, leveraging prior knowledge of the idiomatic expressions for pattern matching(Cook et al., 2008) .

`3) theidioms.` We scraped 1606 of the most common English idioms from `theidioms.com` website using the Beautiful Soup library, resulting in a dataset of 7830 sentences. A few example sentences accompany each idiom. We use the NLTK library for lemmatization and text processing. We used a function to identify positions in sentences where a phrase similar to the idiomatic phrase occurs based on the lemmatized tokens and a similarity threshold. We use a similarity threshold of 0.9, ensuring that even slight variations of the idiomatic phrases are selected and annotated, as the idioms in the example sentences do not maintain the same format across all examples or instances of its usage. We have released a file containing the unfiltered sentences corresponding to particular idiomatic expressions.

`4) formal.` We utilized the EPIE corpus (English Possible Idiomatic Expressions)(Saxena and Paul, 2020), consisting of 25027 sentences. The corpus is divided into Formal and Static idioms, with 3136 sentences containing 358 Formal idioms and 21891 sentences containing 359 Static idioms. Static idioms are expressed using the exact phrase

[2] `https://www.nltk.org/`

in all sentences, whereas formal idioms undergo lexical changes across instances. The token labeling follows the BIO convention with tags B-IDIOM (beginning of PIE), I-IDIOM (continuation of PIE), and O (Non-Idiom token). We merged B-IDIOM and I-IDIOM into one token to match our other datasets and treat this problem as a binary token classification task. We only focus on the formal portion of this dataset as the lexical changes to the expressions address a more robust task.

5) gtrans. We compiled a dataset of 440 sentences using GPT-3.5, featuring 22 English idioms sourced manually from online platforms. Each idiom was paired with 20 example sentences. After translating these idioms to Hindi and then back to English, we observed that Google Translate accurately retained their meanings, demonstrating its understanding of these idioms.

6) gpt+gtrans. We added 440 sentences generated by GPT-3.5 without idiomatic expressions to the gtrans dataset, resulting in a total of 880 sentences. 440 with idiomatic expressions present, and 440 without idioms. Token labeling and annotation followed similar methods as in previous datasets. Additionally the sentences without idioms have all tokens labeled as 0.

7) theidioms 1-1. The dataset, sourced from theidioms.com, contains 1606 idioms (also present in theidioms), each with a single instance, ensuring a 1-1 mapping between sentences and idioms. We labeled tokens using pattern matching and text processing with the NLTK library. This dataset tests the model's generalization by including idioms unseen during training.

5.4. Metrics

Precision, Recall, F1. We calculated precision, recall, and F1-scores for both I and NI classes, presenting them as ordered pairs.
Macro and Weighted Average F1. We calculated macro average as a mean of the values of the ordered pair, and the weighted average considering the relative number of each token in the complete dataset.
Weighted-Averaged Formulae

$$P = \frac{\sum_{i=1}^{N}(TP_i + FP_i) \times P_i}{\sum_{i=1}^{N}(TP_i + FP_i)}$$

$$R = \frac{\sum_{i=1}^{N}(TP_i + FN_i) \times R_i}{\sum_{i=1}^{N}(TP_i + FN_i)}$$

$$F1 - \text{score} = \frac{\sum_{i=1}^{N}(2 \times P_i \times R_i) \times (TP_i + FN_i)}{\sum_{i=1}^{N}(P_i + R_i) \times (TP_i + FN_i)}$$

Where **P**: Precision; $\mathbf{P_i}$: Precision of the i^{th} example; **R**: Recall; $\mathbf{R_i}$: Recall of the i^{th} example; **N**: Number of classes (2 in our case); $\mathbf{TP_i}$: True Positives for class i; $\mathbf{FP_i}$: False Positives for class i; $\mathbf{FN_i}$: False Negatives for class i; $\mathbf{TN_i}$: True Negatives for class i.
Sequence Accuracy. A sentence is only considered correct if all of its constituent tokens are correctly marked. This metric can be considered as a much more stringent metric than normal F1 and accuracy scores (Zeng and Bhat, 2021).

5.5. Results

5.5.1. With Regular Loss

Table 2 shows our results. Our base models utilizing regular binary cross entropy loss display good baseline results, however the results are consistently the lowest across all datasets and experiments compared to using custom loss functions. Our base results on EPIE formal show a large increase in metrics over the results proposed (Gamage et al., 2022). We see an increase of 1.24% in precision, 19.6% in recall and 10.9% in F1-score for the minority idiomatic class.

5.5.2. With Re-translation based Loss

Using re-translation based loss improves precision, recall, and F1 scores over binary cross entropy loss on all the datasets. It leads to large gains on theidioms, theidioms 1-1, formal, gtrans, and gpt&gtrans. This can be explained by the fact that these datasets are characterized by more comprehensive and meaningful sentences compared to MAGPIE and VNC, which often contain phrases and incomplete sentences. We also observe that the translation-based loss exhibits the highest performance on our in-house dataset, gtrans, and this outcome is anticipated, as the expressions included in the dataset primarily rely on the translation model's capacity to grasp the genuine meaning of the idiom in its context and substitute it with a literal phrase conveying the same intended meaning. For the formal corpus, we see further increases of 3.3% in precision, 3.11% in recall and 3.22% in F1-score over our regular loss model. This clearly shows the superiority of translation-based loss function.

5.5.3. With Cohesion based Loss

We conducted an initial study to use cohesion based score to classify sentences into containing an idiom or not. It showed results of around 70% accuracy and varied according to the quality of the datasets. Incorporating it as an objective during training improved the accuracy further on all the datasets compared to regular

		Precision			Recall			F1			
Dataset	Method	Precision	Precision Macro Avg	Precision Weighted Avg	Recall	Recall Macro Avg	Recall Weighted Avg	F1	F1 Macro Average	F1 Weighted Average	Accuracy
MAGPIE	Regular Cross Entropy Loss	[94.1,99.27]	96.68	98.74	[93.64,99.32]	96.48	98.74	[93.87,99.3]	96.58	98.74	98.74
	Translation Retranslation Loss	[93.96,**99.31**]	96.64	98.76	[**93.99**,99.31]	**96.65**	98.76	[93.98,99.31]	96.64	98.76	98.76
	Cohesion based Loss	[94.22,99.28]	96.75	98.76	[93.77,99.34]	96.55	98.76	[93.99,99.31]	96.65	98.76	98.76
	Combination	[**94.5**,99.29]	**96.89**	**98.79**	[93.78,**99.37**]	96.58	**98.8**	[**94.14**,**99.33**]	**96.73**	**98.79**	**98.8**
VNC	Regular Cross Entropy Loss	[97.19,99.64]	98.41	99.43	[96.14,99.74]	97.94	99.43	[96.66,99.69]	98.17	99.43	99.43
	Translation Retranslation Loss	[97.99,**99.81**]	98.9	99.66	[**97.99**,99.81]	98.9	99.66	[97.99,99.81]	98.9	99.66	99.66
	Cohesion based Loss	[98.13,99.76]	98.94	99.62	[97.37,99.83]	98.6	99.62	[97.75,99.79]	98.77	99.62	99.62
	Combination	[**98.45**,**99.81**]	**99.13**	**99.7**	[**97.99**,**99.86**]	**98.92**	**99.7**	[**98.22**,**99.83**]	**99.03**	**99.7**	**99.7**
theidioms	Regular Cross Entropy Loss	[86.61,95.33]	92.07	95.75	[87.37,97.37]	92.36	95.73	[86.98,97.45]	92.21	95.74	95.73
	Translation Retranslation Loss	[91.60,98.68]	95.13	97.52	[93.24,98.33]	95.78	97.5	[92.40,98.50]	95.45	97.51	97.5
	Cohesion based Loss	[91.62,**98.83**]	95.22	**97.65**	[**94.03**,98.32]	**96.17**	**97.62**	**[92.8,98.57]**	**95.69**	**97.63**	**97.62**
	Combination	[**91.76**,98.77]	**95.26**	97.63	[93.73,**98.36**]	96.05	97.61	[92.73,98.56]	95.65	97.61	97.61
formal	Regular Cross Entropy Loss	[90.04,99.18]	94.6	97.89	[95.02,98.29]	96.65	97.82	[92.46,98.73]	95.59	97.84	97.83
	Translation Retranslation Loss	[93.34,99.69]	96.52	98.8	[98.13,98.86]	98.49	98.76	[95.68,99.27]	97.48	98.77	98.75
	Cohesion based Loss	[92.47,**99.75**]	96.11	98.73	[**98.51**,98.69]	**98.6**	98.67	[95.39,99.22]	97.31	98.68	98.67
	Combination	[**93.71**,99.70]	**96.71**	**98.87**	[98.22,**98.92**]	98.57	**98.82**	[**95.92**,**99.31**]	**97.61**	**98.84**	**98.83**
gtrans	Regular Cross Entropy Loss	[85.93,93.87]	89.9	92.38	[72.39,97.27]	84.83	92.61	[78.54,95.53]	87.04	92.36	92.61
	Translation Retranslation Loss	[**86.94**,**96.71**]	**91.83**	**94.89**	[85.68,**97.03**]	**91.36**	**94.91**	[**86.30**,**96.87**]	**91.59**	**94.9**	**94.91**
	Cohesion based Loss	[86.76,**96.71**]	91.74	94.85	[**85.69**,96.99]	91.33	94.87	[86.21,96.85]	91.53	94.86	94.87
	Combination	[86.86,96.58]	91.72	94.76	[85.07,**97.03**]	91.05	94.79	[85.94,96.80]	91.38	94.77	94.79
gpt&gtrans	Regular Cross Entropy Loss	[80.4,97.84]	89.12	96.09	[80.79,97.78]	89.29	96.06	[80.53,97.81]	89.17	96.07	96.07
	Translation Retranslation Loss	[83.91,98.85]	**91.38**	97.34	[89.83,98.06]	93.94	97.23	[86.74,98.45]	92.59	97.27	97.23
	Cohesion based Loss	[83.05,**99.02**]	91.03	**97.41**	[**91.37**,97.91]	**94.62**	**97.25**	[**86.99**,**98.46**]	**92.73**	**97.3**	**97.25**
	Combination	[**83.97**,98.83]	91.4	97.33	[89.64,**98.08**]	93.86	97.23	[86.70,98.45]	92.58	97.26	97.22
theidioms 1-1	Regular Cross Entropy Loss	[66.53,92.80]	79.67	88.91	[57.58,94.97]	76.27	89.44	[61.73,93.87]	77.8	89.12	89.44
	Translation Retranslation Loss	[72.47,93.24]	82.85	90.17	[59.90,**96.05**]	77.97	90.7	[65.58,94.63]	80.1	90.33	90.56
	Cohesion based Loss	[71.88,**93.49**]	82.69	90.3	[**61.59**,95.82]	**78.71**	90.75	[**66.37**,94.64]	80.18	90.45	90.75
	Combination	[**72.84**,93.40]	**83.11**	**90.36**	[60.89,**96.05**]	78.47	**90.85**	[66.31,**94.71**]	**80.51**	**90.51**	**90.85**

Table 2: Results of applying idiom-based custom loss function on several datasets

binary cross entropy loss. As observed for translation-based loss, it leads to large gains on the theidioms, theidioms 1-1, formal, gtrans, and gpt&gtrans, and performs the best on the theidioms and gpt&gtrans datasets because these datasets contain sentences which are more complete than MAGPIE and VNC. For formal corpus, we see further increases of 2.43% in precision, 3.49% in recall and 2.93% in F1-score over our regular loss model. This observation aligns perfectly with the fundamental concept of our metric. It underscores that idioms embedded within highly cohesive sentences are more readily identifiable as being idiomatic usages of those phrases.

5.5.4. With combination of losses

Using a combination of both losses improves the accuracy values on MAGPIE, VNC, formal, and theidioms 1-1 and is very close to the accuracies of translation-based or cohesion-based loss functions for other datasets. In formal corpus, we observe notable improvements: precision increases by 3.67%, recall by 3.2%, and F1-score by 3.46% compared to our regular loss model. These discoveries validate the efficacy of utilizing both semantic cohesion and dissimilarity of idiomatic phrases within their contextual environments for our task. Instances penalized by both metrics typically represent confidently idiomatic expressions, which the model should strive to accurately classify.

5.5.5. Cross-domain performance across datasets

We trained our models on one dataset and tested them on another to measure the generalization ca-

		Precision			Recall			F1			
Train, Test	Method	Precision	Precision Macro Avg	Precision Weighted Avg	Recall	Recall Macro Avg	Recall Weighted Avg	F1	F1 Macro Average	F1 Weighted Average	Accuracy
`theidioms, gtrans`	Regular Cross Entropy Loss	[84.73,96.41]	90.39	94.11	[84.78,96.29]	90.54	94.1	[84.57,96.35]	90.46	94.11	94.1
	Translation Retranslation Loss	[**89.3**,98.21]	93.76	96.51	[92.46,**97.39**]	94.93	96.45	[90.85,97.8]	94.33	96.48	96.45
	Cohesion based Loss	[**89.3**,**98.39**]	**93.84**	**96.65**	[**93.21**,97.37]	**95.29**	**96.57**	[**91.21**,**97.87**]	**94.54**	**96.6**	**96.57**
	Combination	[89.20,97.97]	93.59	96.3	[91.42,**97.39**]	94.4	96.25	[90.28,97.68]	93.98	96.27	96.25

Table 3: Results showing transfer capabilities of our models. The model is trained on `theidioms` and tested on `gtrans`.

pabilities of the model and how our methodology may improve this capability. We trained the model on the `theidioms` dataset and tested on `gtrans` dataset. Table 3 shows the result. Our custom loss function based approach showcases impressive transfer capabilities.

5.5.6. Comparison with DISC

We compared our models with DISC (Zeng and Bhat, 2021), a state-of-the-art approach for idiom token classification. We refer to the accuracy values reported in the paper to compare our technique with theirs. We kept the same train-test split for MAGPIE and VNC dataset. It should also be noted that DISC was trained for 600 epochs while our models were trained for only 5 epochs. Table 4 compares the **sequence accuracies** of DISC and our model. Sequence accuracy is considered as a better metric to capture the performance of such models (Zeng and Bhat, 2021). It is clear that our model outperforms DISC in sequence accuracy. This can be explained by our model's capabilities in distinguishing between the literal and figurative idiomatic usages, possible through custom loss function training.

Dataset	Method	Sequence Accuracy
MAGPIE	Regular Cross Entropy Loss	90.19
	Translation Retranslation Loss	91.31
	Cohesion based Loss	91.46
	Combination	**91.51**
	DISC[3]	87.47
VNC	Regular Cross Entropy Loss	93.75
	Translation Retranslation Loss	**96.88**
	Cohesion based Loss	**96.88**
	Combination	**96.88**
	DISC	93.31

Table 4: Comparing DISC, a state-of-the-art idiom detection model with our technique on 2 datasets

6. Discussion

When we consider the examples where the DISC approach is making incorrect predictions, for instance - "Dragons can lie for dark centuries brooding over their treasures, bedding down on frozen flames that will never see the light of day." The DISC approach incorrectly predicts only a portion of the complete expression - "see the light of day" as idiomatic, whereas our model correctly identifies the entire expression. Similarly for - "Given a method, we can avoid mistaken ideas which, confirmed by the authority of the past, have taken deep root, like weeds in men's minds." where the DISC model predicts "weeds in men's minds" as the idiomatic expression with the correct instance being "taken deep root". Our models do not falter in this case and predict all tokens for this example correctly.

In instances where the cohesion-based approach outperforms combined approaches, it is noteworthy that the Multi-Word Expressions (MWEs) are not consistently translated as expected. Consequently, the incorporation of the translation score tends to diminish overall performance. On the other hand, the translation-only model demonstrates an ability to enhance results compared to the baseline, as it successfully captures anticipated translations for certain expressions, contributing to improved overall performance.

We manually analyze the different errors that our models make on the VNC and EPIE formal datasets to gain insights into the idiom identification abilities and shortcomings in Table 5. We have categorized the errors into 5 major cases and we present examples of each type. Case 1 is where the correct idiomatic expression is identified fully but an alternate expression has also been tagged as idiomatic. This can be thought of as a limitation of the datasets rather than that of our models, as our datasets label at most one expression as idiomatic in each sentence. The second case is where an alternate expression is labeled. The reasoning for this is similar to the previous case as there may be multiple expressions that could possibly be idiomatic and our model is identifying one of them. In the third case, our model correctly identifies the idiom but also tags words surrounding the idiom. This can be ascribed to the alterna-

Error Type	Sentence with PIE	Prediction
Multiple Expressions Predicted	I then walked across to the photographers and lost my temper and then *lost my head*.	lost my temper , lost my head
Alternate expression detected	Cantona will have to *kick his heels* on the sidelines if the manager had his way.	had his way
Extra tokens surrounding expression	Julia had her *attention caught* by the commotion.	attention caught by
Partial	His blistering turn of speed and attitude *made him an instant hit* with the fans.	hit
Predicting Nothing	Everyone talks about *hitting a wall* at the 24 mile mark.	Empty String

Table 5: Different error types along with examples and the incorrect prediction. The ground truth values have been colored blue in sentences.

tive labeling of the identical expression in different occurrences. The fourth case ”Partial”, constitutes instances where only a segment of the idiomatic expression is identified, with the specific localization of the entire idiom boundary remaining imprecise. The last error category involves the absence of predictions when the model fails to recognize idiomatic usage, even when it is present. The effectiveness of our model is contingent upon the caliber of annotation and various other external factors.

7. Future Work

The latest advancements in Natural Language Processing (NLP) have led to the extensive utilization of a range of transformer-based models. We can adjust our own loss functions to refine different architectures effectively. We can create an intuitive and efficient tool utilizing these fine-tuned models to detect an idiom in a given sentence. This tool should offer a straightforward and accessible experience for a broad range of users, with minimal technical expertise required. To continuously improve the overall performance of our models, we can systematically address each identified error category. This might involve analyzing error patterns and refining the fine-tuning process accordingly.

8. Acknowledgements

We are grateful to the anonymous reviewers for their insightful comments which substantially improved this manuscript. This work was performed using *Sharanga*, the high performance computing cluster at the BITS Pilani Hyderabad Campus.

9. Bibliographic References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Rhys Biddle, Aditya Joshi, Shaowu Liu, Cecile Paris, and Guandong Xu. 2020. Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter. In *Proceedings of The Web Conference 2020*, WWW ’20, page 1217–1227, New York, NY, USA. Association for Computing Machinery.

J Briskilal and CN Subalalitha. 2022. Classification of Idiomatic Sentences Using AWD-LSTM. In *Expert Clouds and Applications: Proceedings of ICOECA 2021*, pages 113–124. Springer.

BNC Consortium. 2007. British National Corpus, XML edition. Oxford Text Archive.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22. Citeseer.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 337–344.

Gihan Gamage, Daswin De Silva, Achini Adikari, and Damminda Alahakoon. 2022. A BERT-based Idiom Detection Model. In *2022 15th International Conference on Human System Interaction (HSI)*, pages 1–5.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Callum Hughes, Maxim Filimonov, Alison Wray, and Irena Spasić. 2021. Leaving no stone unturned: flexible retrieval of idiomatic expressions from a large text corpus. *Machine Learning and Knowledge Extraction*, 3(1):263–283.

Sami Itkonen, Jörg Tiedemann, and Mathias Creutz. 2022. Helsinki-NLP at SemEval-2022 Task 2: A Feature-Based Approach to Multilingual Idiomaticity Detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 122–134.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3230–3236. AAAI Press.

Changsheng Liu and Rebecca Hwa. 2019. A generalized idiom usage recognition model based on semantic compatibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 01, pages 6738–6745.

Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuanjing Huang. 2017. Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213, Copenhagen, Denmark. Association for Computational Linguistics.

Vasudevan Nedumpozhimana, Filip Klubička, and John D Kelleher. 2022. Shapley idioms: Analysing BERT sentence embeddings for general idiom token identification. *Frontiers in Artificial Intelligence*, 5:813967.

Jing Peng and Anna Feldman. 2017. Automatic idiom recognition with word embeddings. In *Information Management and Big Data: Second Annual International Symposium, SIMBig 2015, Cusco, Peru, September 2-4, 2015, and Third Annual International Symposium, SIMBig 2016, Cusco, Peru, September 1-3, 2016, Revised Selected Papers 2*, pages 17–29. Springer.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. "Classifying Idiomatic and Literal Expressions Using Topic Models and Intensity of Emotions". In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Priyanka and R.M.K. Sinha. 2014. A system for identification of idioms in hindi. In *2014 Seventh International Conference on Contemporary Computing (IC3)*, pages 467–472.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th international joint conference on natural language processing*, pages 210–218.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom Token Classification using Sentential Distributed Semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. EPIE Dataset: A Corpus For Possible Idiomatic Expressions.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010.

Patrizia Tabossi, Rachele Fanari, and Kinou Wolf. 2008. Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2):313.

Patrizia Tabossi, Rachele Fanari, and Kinou Wolf. 2009. Why are idioms recognized fast? *Memory & cognition*, 37:529–540.

Minghuan Tan and Jing Jiang. 2021. Does BERT Understand Idioms? A Probing-Based Empirical Study of BERT Encodings of Idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. Id10m: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726.

Rakesh Verma and Vasanthi Vuppuluri. 2015. A new approach for idiom identification using meanings and the web. In *Proceedings of the international conference recent advances in natural language processing*, pages 681–687.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Ziheng Zeng and Suma Bhat. 2022. Getting BART to Ride the Idiomatic Train: Learning to Represent Idiomatic Expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Ziheng Zeng, Kellen Tan Cheng, Srihari Venkat Nanniyur, Jianing Zhou, and Suma Bhat. 2023. IEKG: A Commonsense Knowledge Graph for Idiomatic Expressions. *arXiv preprint arXiv:2312.06053*.

Ad Hoc Compounds for Stance Detection

Qi Yu[1,2], **Fabian Schlotterbeck**[3], **Hening Wang**[3], **Naomi Reichmann**[1],
Britta Stolterfoht[3], **Regine Eckardt**[1,2], **Miriam Butt**[1,2]
[1]Department of Linguistics, University of Konstanz
[2]Cluster of Excellence "The Politics of Inequality", University of Konstanz
[3]Department of Modern Languages, University of Tübingen
firstname.lastname@{uni-konstanz, uni-tuebingen}.de

Abstract

In this paper we focus on a subclass of multi-word expressions, namely compound formation in German. The automatic detection of compounds is a known problem and we argue that its resolution should be given more urgency in light of a new role we uncovered with respect to ad hoc compound formation: the systematic expression of attitudinal meaning and its potential importance for the down-stream NLP task of stance detection. We demonstrate that ad hoc compounds in German indeed systematically express attitudinal meaning by adducing corpus linguistic and psycholinguistic experimental data. However, an investigation of state-of-the-art dependency parsers and Universal Dependency treebanks shows that German compounds are parsed and annotated very unevenly, so that currently one cannot reliably identify or access ad hoc compounds with attitudinal meaning in texts. Moreover, we report initial experiments with large language models underlining the challenges in capturing attitudinal meanings conveyed by ad hoc compounds. We consequently suggest a systematized way of annotating (and thereby also parsing) ad hoc compounds that is based on positive experiences from within the multilingual ParGram grammar development effort.

Keywords: ad hoc compounds, attitudinal meaning, stance detection, German, universal dependencies, psycholinguistic validation, large language models

1. Introduction

The automatic detection of compounds is known to be a difficult problem for Natural Language Processing (NLP) (Constant et al., 2017; Baldwin and Kim, 2010), particularly in a language like German which uses compounding as a central strategy for novel word formation. In this paper we present research showing that novel, ad hoc compound formations in German can be used strategically to convey attitudinal meaning, thus making them an interesting area of research from the overall perspective of stance detection (Mohammad et al., 2016; Schiller et al., 2021) and adding urgency to finding reliable ways of automatically detecting compounds, and in particular, novel compound formations in a language. We adduce evidence that combines insights from theoretical linguistic analysis, corpus linguistic investigations and psycholinguistic experimentation to show that a subset of ad hoc compounds in German, termed *enigmatic compounds* (ECs; Wildgen, 1981) are indeed systematically used to convey attitudinal meaning and are therefore of inherent interest for the NLP task of stance detection.

The types of compounds falling under the rubric of ECs are illustrated in (1)–(3). We noted the use of such compounds for the expression of stance as part of a larger project investigating the framing of politically charged issues across several German newspapers. We have marked the extra expressive meaning carried by these ad hoc compound formations as attitudinal meaning (AM) in the examples.

(1) Flüchtlinge wollen Österreich meiden und
refugees want Austria avoid and
lieber in Merkel-Land einreisen.
rather in Merkel Country travel.into
'Refugees want to avoid Austria and instead enter Merkel-Country.'
AM: The German refugee crisis is Merkel's fault.
(SOURCE: Facebook)

(2) Jede 5. China-Maske ist unbrauchbar
every fifth China mask is unusable
'Every fifth China-mask is unusable'
AM: China is notorious for low-quality products.
(SOURCE: BILD, 2020-05-03)

(3) Neue Stelle für Kopftuch-Praktikantin
new position for hijab intern
'New position for hijab-intern'
AM: Religious practices of Muslims often cause trouble for others.
(SOURCE: BILD, 2016-08-25)

Intended but deliberately masked meanings of speakers such as the AMs above are known to play a crucial role in political communication (Beaver and Stanley, 2018). Our data indicate that ECs are a useful rhetorical device for speakers/authors to implicitly convey attitudinal meaning. In particular, we observed that ECs can be employed as so-called *dog-whistles* (Henderson and McCready, 2019), whereby their use – at least for a certain time span – speaks to a certain subgroup and con-

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024) @LREC-COLING-2024, pages 231–242. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

veys a meaning that is on the surface rather vague, but decodable as to its hidden meaning by that subgroup. This seems particularly interesting, as ad hoc compounds are instances of innovated language and thus, dog whistles and pejorative uses in expressing attitudinal meaning clearly cannot rest on conventional lexical meanings alone. This makes an automatized stance detection task challenging yet interesting.

We consequently examine how compounds are currently treated in available dependency parsers and Universal Dependencies (UD) treebanks (de Marneffe et al., 2021; Nivre et al., 2016) for German. We find that the current treatment is uneven. We also explored the potentially greater capabilities of current large language models (LLMs) with respect to detecting attitudinal meaning in ECs, but found that while the results from LLMs may provide an explanation for substantial variation in our experimental data, they do not easily capture the effect of our experimental manipulation involving ECs. We therefore provide suggestions for a systematic UD annotation for compounds that is based on the multilingual ParGram grammar development experience (Butt et al., 1999; Sulger et al., 2013) so as to allow for a more successful learning.

This paper is structured as follows: in section 2 we provide some background on the current state-of-the-art. We follow this in section 3.1 with the results of a corpus study of three different newspapers, which yielded indications that more conservative leaning newspapers used ad hoc compounds to trigger attitudinal meaning more than other newspapers. However, our results are most robust for the conservative tabloid *BILD*, which is also known for an editorial policy that prefers the use of pictures coupled with short, expressive texts. The greater use of ad hoc compounds could also therefore just be a matter of newspaper writing style. To test the perception of attitudinal meaning in compounds, we therefore designed and executed an experiment that sought to establish the stance triggering effect of ECs using psycholinguistic methods. This is described in section 3.3, following a discussion of how the semantics of ECs are hypothesized to come about in section 3.2. In section 4, we report on our attempts to use current LLMs to simulate our experimental results. We did not find any indication that these models can capture the central contrasts observed in the experimental outcomes. In section 5, we combine the insights from the corpus and psycholinguistic results to formulate recommendations for the systematic annotation of compounds in corpora. Section 6 concludes the paper.

2. Background

2.1. Evaluative Language

Evaluative language is of interest for a range of NLP tasks, perhaps currently most prominent among the sentiment analysis (Pang and Lee, 2008; Taboada et al., 2011), but also hate speech detection (Davidson et al., 2017) and stance detection (Mohammad et al., 2016; Schiller et al., 2021). Sentiment analysis and stance detection are closely related tasks but differ in their overall goals. Sentiment analysis is concerned with identifying whether a given text, sentence or passage overall can be classified as being positive, negative or neutral. This has generally involved a bag-of-words approach, where the internal structure of the text is not considered and the target has generally been reviews or statements about movies, books, objects or persons. More recently, approaches to sentiment analysis have become more nuanced in that the classification aims at *aspect based* (what aspect is the sentiment targeted at, e.g., the acting or the plot?) or *target based* (what is the precise target of the sentiment, e.g. an iPhone or the ear phones that came with the iPhone?) sentiment analysis (Alturayeif et al., 2023).

Stance detection is informed by the Stance Triangle defined by Du Bois (2007), by which the author of a text is taken to want to influence or align the recipient/reader of the text with his/her beliefs. The difference between sentiment analysis and stance detection is that in sentiment analysis the object of study are texts expressing a given sentiment, prototypically reviews. In these the author articulates their opinion to an audience, but is not necessarily seeking to align the audience with their own views. Given that our overall interest lies in determining how issues are framed (Chong and Druckman, 2007), we are interested in stance detection as a subtask for determining the overall framing of a narrative or text. As far as we have been able to determine, no previous work on stance detection has attempted to include information from compounds in a focused manner, though Li and Caragea (2019) note as part of their stance detection error analysis that it would be useful to separate the individual components used in hashtags such as as *#VoteGOP* or *#NoHilary*, as found in the SemEval-2016 dataset developed specifically for stance detection (Mohammad et al., 2016).

Stance detection includes identifying instances of subjective language (Wiebe et al., 2004). Subjective language can be detected on the basis of linguistically informed lexicon and/or construction based information (Biber and Finegan, 1989; Biber and Conrad, 2019; Taboada et al., 2011), or it can be detected by machine learning on the basis of annotated data (Alturayeif et al., 2023). Our data is German, for which an automatic annotation tool for subjective language already exists (El-Assady et al., 2016, 2019). This tool provides POS-tagging and syntactic parsing of a given text along with a systematic identification of linguistic cues for subjective language such as the annotation of various modals or German discourse particles (Zimmermann, 2011). However, the tool does not include a facility for the automatic detection of ECs.

2.2. Annotation and Automatized Detection of Compounds

The compounds in (1)–(3) each contain a hyphen. However, German compounds generally do not contain a hyphen. One could hypothesize that ad hoc compounds in particular are marked with a hyphen, but our data also contains instances of ad hoc formations such as *Asylprügler* 'asylum beater' and *Migrantenschreck* 'migrant scare' that have been written without a hyphen. Nevertheless, the inclusion of a hyphen provides a potentially important clue for the automatic identification of at least a subset of compounds and one that could be picked up on easily. In surveying existing dependency parsers and treebanks annotated according to the Univeral Dependencies (UD) scheme (de Marneffe and Manning, 2008; de Marneffe et al., 2021; Nivre et al., 2016), we found that only the Stanza toolkit (Qi et al., 2020) could reliably identify German compounds characterized by a hyphen. The sample of other dependency parsers for German that we tried were not reliable in the identification of compounds, with most merely labeling them with the POS-tag of NN for common nouns, as shown in Figure 1 for the Mate parser (Björkelund et al., 2010),[1] where both of the compounds *Flüchtlingsorganisation* 'refugee organisation' and *Asyl-Verschärfungen* 'asylum restrictions' are tagged as NN. The same is true for spaCy,[2] ParZu (Sennrich et al., 2009)[3] and a German dependency parser[4] based on the MaltParser framework[5], as well as the very high quality morphological analyzer SMOR (Schmid et al., 2004). An investigation of UD treebanks for German collected at the INESS website[6] yielded much the same result. See also the reports and conclusions in Baldwin et al. (2023).

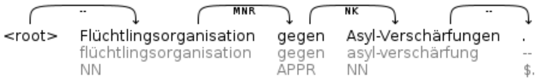


Figure 1: Sample Mate parse

A morphological analyzer can be integrated as part of a dependency parser and so we set out to test SMOR for our purposes. We worked with this system because it has been designed especially to deal with the productive word formation possibilities in German, including ad hoc compounds occurring in newspaper texts. However, a pilot study with respect to our data showed that while SMOR is indeed able to identify ad hoc compounds successfully, the uneven nature of the overall results means that quite a bit of manual postprocessing would be required to obtain a useable data set. For example, the ad hoc compound *Pegida-Anhänger* 'Pegida follower/supporter' could not be analyzed at all, while the lexically established word *Bezirksamt* 'district office' was incorrectly analyzed. Instead of the correct split into the morphemes *Bezirk+s+amt* (the *s* is a so-called linking element that appears for phonological reasons), the word was split into *Bezirk+Samt* 'district velvet' as one of the three most likely results.

Thus, the challenges posed by automatic compound detection (Constant et al., 2017; Baldwin and Kim, 2010) continue to be a problem, and one that – we argue – gains more urgency given our findings. Given that ECs express attitudinal meaning and as such can provide an important linguistic cue for stance detection, search for these cues should be operationalized.

3. Enigmatic Compounds

In this section, we combine results from a corpus linguistic study and a psycholinguistic experiment to show that ECs can be used systematically to express attitudinal meaning. We first present results from a corpus study that demonstrates a systematic use of ECs to express a negative stance in newspapers (section 3.1). We then discuss how ad hoc compounds invite such attitudinal meaning from a theoretical linguistic aspect (section 3.2), and report a psycholinguistic experiment (section 3.3) to confirm that ECs are indeed a systematic part of language use. All data and code resulting from this work are publicly available at: `https://github.com/qi-yu/enigmatic-compounds`.

3.1. Corpus Study

Our corpus study was conducted as part of a larger investigation into the framing of the Syrian refugee crisis by German newspapers in the time span of

[1] `https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools/`

[2] `https://spacy.io/`

[3] `https://github.com/rsennrich/ParZu`

[4] `https://pub.cl.uzh.ch/users/siclemat/lehre/ecl1/ud-de-hunpos-maltparser/html/`

[5] `http://www.maltparser.org/`

[6] `https://clarino.uib.no/iness-prod/treebanks#`

2014–2018. We chose the three German newspapers with the highest circulation rates (IVW, 2023): *BILD*, *Frankfurter Allgemeine Zeitung* (FAZ) and *Süddeutsche Zeitung* (SZ). These three newspapers cover a representative range of political leanings within the German media landscape, with *BILD* being the most conservative on the political spectrum, the SZ the most left leaning, and the FAZ also leaning towards the conservative end. Moreover, they also build a diverse sample of different styles, with BILD characterized as a tabloid newspaper whereas FAZ and SZ contain high quality, in-depth reporting. Examples (4)–(6) illustrate the different styles: they are headlines from articles reporting on the same event and published around the same time.

(4) **BILD**, 2014-09-29:
Folter-Skandal in deutschen Asylbewerberheimen: "Die Wachleute schlagen und treten uns"
'Torture-scandal in German asylum seekers' accommodations: "The guards beat and kick us"'

(5) **FAZ**, 2014-09-30:
Misshandlung von Asylbewerbern: Sicherheitsleute werden überprüft
'Mistreatment of asylum seekers: security guards undergo checks'

(6) **SZ**, 2014-09-30:
Ermittlungen nach Misshandlungsverdacht in drei Flüchtlingsheimen
'Investigations into suspected mistreatment in three refugee accommodations'

As part of this investigation, we noticed that compounds seemed to be used to express a negative stance towards refugees and the handling of the crisis by the government (see, e.g., *Folter-Skandal* 'torture-scandal' in (4)). A more in-depth investigation of this phenomenon was hampered by the difficulty of automatically detecting compounds. We therefore decided to experiment with training a language model on the basis of annotated data. The best performing model was a logistic regression model that resulted in a value of 0.68 for F1.

Given these unsatisfactory results, we asked ourselves whether it was indeed necessary to detect these compounds. As we report on in the following sections, the result of our investigations has established that ECs indeed have the potential for providing important information for stance detection. Efforts should be redoubled so as to be able to operationalize ECs for stance detection.

Our data set consisted of a total of 23,889 articles. Given the necessity for manual annotation of the compounds (since automatic detection is a challenge), we considered only the articles' headlines for our study. We manually identified 19,353 referential/neutral ad hoc compounds and 828 ECs in these headlines. We structured our resulting data set into pieces of information as follows: the target compound, the sentence in which it appeared, the year it was released, the newspaper source, and the annotation (`0` = referential, `1` = enigmatic). We categorized the compounds as enigmatic if they met the following two criteria:

(i) the compound carries an attitudinal meaning;

(ii) the compound is an innovative, ad hoc formation and is thus not established in a recognized dictionary or lexicon of German.

To validate the application of criterion (ii), the German dictionary *Duden*[7] as well as the online dictionary *Digitales Wörterbuch der deutschen Sprache*[8] were consulted. For instance, based on these criteria, the compound *Karajan-Schüler* 'Karajan student' was defined as referential (neutral), as it does not seem to express an additional evaluative meaning; only its literal meaning is being transmitted. In contrast, the compound *Flüchtlings-Tsunami* 'refugee tsunami' was categorized as enigmatic, as it does not only refer to a large amount of refugees, but it also carries an additional AM to the effect that refugees are overwhelming the transit and host countries.

Our overall results of the annotation per newspaper are given in Table 1. They show that *BILD* uses by far the most ECs. We furthermore sampled the top most ECs per newspaper per year and found that *BILD* predominantly used these in contexts of discussing security or issues of criminality, whereas the FAZ and the SZ placed a greater emphasis on problems of capacity and the rights of individual refugees. For example, with compounds such as *Asylprügler* 'asylum beater', *Migrantenschreck* 'migrant scare', and *Amok-Afrikaner* 'amok African', *BILD* focuses on criminality related to the refugees in Germany through the use of ECs. This is in line with the hostile reporting style previously observed for tabloid newspapers (see Innes, 2010; Kleinsteuber and Thomass, 2007).

Newspaper	#Enigmatic	#Neutral
BILD	726	10,059
FAZ	58	5,525
SZ	44	3769

Table 1: Total number of enigmatic and neutral compounds in newspaper headlines.

Whether or not the ECs are employed as attention-getters as part of *BILD*'s sensationalist

[7] https://www.duden.de

[8] https://www.dwds.de

writing style (see Greussing and Boomgaarden, 2017) becomes irrelevant in the face of their extensive use by *BILD* in combination with the negative attitudinal meanings triggered by these ECs: they are a significant contributing factor to the overall articulated stance towards a topic.

3.2. Compound Meaning

Compounds have a range of interpretational possibilities because their meanings are not compositional. Earlier theoretical linguistic studies on compound meaning share the common assumption that there is some covert, meaning-decisive *semantic relation* $\mathcal{R}$ between the constituents of a compound:

(7) Let C_1C_2 be a compound where $[\![C_1]\!] = m_1$ and $[\![C_2]\!] = m_2$.
Then: $[\![C_1C_2]\!] = \mathcal{R}(m_1, m_2)$

Levi (1978) and Fanselow (1981) propose taxonomies of semantic relations that play a role in *ad hoc* compound interpretation, and Meyer (1993), Ryder (1994) and Benczes (2009) propose different assumptions on how the semantic relations in (7) are derived. In the simplest case, *ad hoc* compounds serve as abbreviations for phrases, as in *Karajan-Schüler* 'Karajan Student' which is equivalent to *Schüler von Karajan* 'student of Karajan'. In (1)-(3); however, there is clearly an attitudinal meaning, an extra meaning dimension that is not found in the equivalent non-compound phrase. Consider, for example, *China-Maske* 'China mask' in the context in (2): it has a negative attitudinal meaning that is not conveyed by the compositional alternative phrase *chinesische Maske* 'Chinese mask'.

Sassoon (2011) opens an avenue towards an explanation of attitudinal enrichment in ECs. The author summarizes comparative studies in the conceptual structure of nouns and adjectives: nouns denote similarity-based concepts with a prototype structure (Murphy, 2002), whereas adjectives denote rule-based properties (Kennedy, 1999). The distinction is backed up by converging evidence from neurolinguistics, patholinguistics and language acquisition. Sassoon's proposal predicts that the modifier in ECs (*China-* in *China-Maske*) contributes to a similarity-based concept. This happens, plausibly, by adding a further dimension in which exemplars must match the prototype. Specifically, similarity-based categorization rests on prototypical values that can be attributed to this dimension. In our example the similarity-based categorization invites a comparison to typical 'products from China', which provides a hook for the accommodation of an interpretation including negative expectations about products from China. The corresponding adjective in a phrasal alternative ('Chinese mask'), in contrast, adds a simple categorical property 'be Chinese' (yes/no). Sassoon thus predicts that the processing of modifiers does not trigger novel stereotypes and should not provide an entry-point for attitudinal meaning.

We were interested in this prediction as it also provides a systematic way of testing whether the attitudinal meaning associated with ECs we found as part of the corpus study in section 3.1 is a general, systematic part of language use or whether it is perhaps attributable to the particular corpus. If the attitudinal meaning associated with ECs is found to be a systematic part of language, it provides another argument for taking ECs seriously as part of the overall task of stance detection. We describe the psycholinguistic experiment we set up to test Sassoon's prediction in section 3.3.

3.3. Experiment

3.3.1. Methods

Materials and Design We manually selected 21 text snippets from newspapers and social media which contain ECs along the lines of (1)–(3) that trigger negative AM according to our own intuitions. We restricted ourselves to negative AMs in our experiment as these were more prevalent in the corpus study. To test for the AM-triggering effects of ECs, three variants were created from each snippet. Table 2 provides examples of such snippets (translated into English). The three variants were:

(i) COMPOUND: original text snippet with the EC.

(ii) PHRASAL: EC substituted by a corresponding phrasal construction.

(iii) NEUTRAL: EC substituted by a corresponding noun that is attitudinally neutral.

The PHRASAL condition controls for truth-conditional information, as it conveys the same truth-conditional information as the COMPOUND condition but in a pragmatically unmarked phrasal expression, not an ad hoc compound. The condition NEUTRAL is intended as a baseline: though there is no stylistic difference in terms of innovative language use between PHRASAL and NEUTRAL, these two conditions differ in their information load, as the modifier part of the PHRASAL (and COMPOUND) condition provides extra information that is not necessary for reference resolution but can be inferred from the prejacent context (see Table 2). Comparing the PHRASAL and the NEUTRAL condition thus allows us to examine whether the addressees' perception of the attitudinal strength is affected by such additional but in principle unnecessary information while keeping the style constant. With these three conditions, we test the following two hypotheses:

(i) COMPOUND VS. PHRASAL (different style, same information load): compounding amplifies the perceived attitudinal strength;

(ii) PHRASAL VS. NEUTRAL (same style, different information load): the additional information that is not necessary for reference resolution amplifies the perceived attitudinal strength.

The items were distributed over 3 lists using a Latin square. 24 stylistically similar text snippets were added to each list as fillers. For each item, participants rated its attitudinal strength by answering question (8) on a 7-point Likert-scale.

(8) *How does the author talk about ________?*
1=positive ○1 ○2 ○3 ○4 ○5 ○6 ○7 *7=negative*

As our overall interest is in the framing of politically charged discourse, we also collected the political leaning of each participant by asking question (9) at the end of the experiment. This allows us to further control whether participants' perception of attitudinal strength is affected by their political leaning:

(9) *In politics, people often use "left" and "right" to denote political leanings. Where would you place your own political leaning?*
1=left ○1 ○2 ○3 ○4 ○5 ○6 ○7 *7=right*

Participants The participant recruitment and data collection was carried out online via Prolific.[9] 212 German native speakers, identified through Prolific's demographic prescreening function, took part in the study (103 female, 102 male, 7 other genders; mean age = 26.52 years, $SD = 8.10$ years). The experiment was carried out anonymously and voluntarily. Each participant received a compensation of £8.50 per hour, a fair rate suggested by Prolific.

3.3.2. Results

Figure 2 shows the rating distributions of each condition. Overall, all items were rated rather negatively, with more negative ratings for COMPOUND than PHRASAL and PHRASAL than NEUTRAL conditions. We fitted a *cumulative link model* (CLM) with random effects using the R package *ordinal* (Christensen, 2018) to test these differences statistically. CLM is a variant of logistic regression generalized to multinomial ordinal dependent variables. A CLM models the probability, $P(Y \leq j)$, that an ordinal response variable Y is less than or equal to a specific category $j \in \{1, \dots, J\}$ ($J \geq 2$) according to the equation below, where θ_j is the intercept of level j, $\mathbf{x}$ is a vector of predictors, and $\boldsymbol{\beta}_j$ is a vector of coefficients:

$$\text{logit}(P(Y \leq j)) = \log \frac{P(Y \leq j)}{P(Y > j)} = \theta_j - \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}$$

In our initial model, we predicted participants' *ratings* using *condition* and participants' *political leaning* as well as their interactions. For the predictor *condition*, PHRASAL is set as reference level (cf. hypotheses above). For the predictor *political leaning*, we mapped the seven original levels (see (9) above) to three aggregated levels in order to ease the model interpretation: 1-3 = LEFT, 4 = NEUTRAL, 5-7 = RIGHT. We used dummy encoding to code the three levels. Random intercepts and random slopes were fitted for *items* and *participants*, as likelihood ratio tests showed that they improved the model fit. In a following model selection step based on likelihood ratio tests, the predictor *political leaning* and the interaction term were removed as they were not significant in improving the model fit (likelihood ratio test without interaction: $\chi^2(2) = 0.384$, $p = 0.826$; likelihood ratio test with interaction: $\chi^2(6) = 2.004$, $p = 0.919$).

Our final model showed a significant difference between COMPOUND and the reference level PHRASAL. Compared to PHRASAL, COMPOUND led to a significant decrease in the logit of ratings in lower (i.e., more positive) categories (COMPOUND VS. PHRASAL: $\beta = 0.526, SE = 0.152, p < 0.001$). No significant difference between NEUTRAL and PHRASAL was found (NEUTRAL VS. PHRASAL: $\beta = -0.272, SE = 0.176, p = 0.123$).

3.3.3. Discussion

The result of our experiment with a large population is in line with the corpus study. The significant decrease of the likelihood of positive ratings indicates that the authors' negative attitudes are perceived as more pronounced when ECs are used instead of the PHRASAL counterpart. The difference in information load between PHRASAL and NEUTRAL condition did not show significant influence on the participants' perception of attitudinal strength. Furthermore, the non-significant effect of *political leaning* as well as the non-significant interaction between *political leaning* and *condition* show that the increased perception of attitudinal meaning in ECs is general part of how language works, rather than being domain or population specific.

4. Simulations with Large Language Models (LLMs)

Recent advances of LLMs have underscored their remarkable utility across a wide variety of NLP

[9] https://www.prolific.co

COMPOUND	PHRASAL	NEUTRAL
The federal government purchased more than 108 million masks from China for German clinics and medical practices. However, about 10 percent of these **China-masks** are unusable for medical purposes.	The federal government purchased more than 108 million masks from China for German clinics and medical practices. However, about 10 percent of these **Chinese masks** are unusable for medical purposes.	The federal government purchased more than 108 million masks from China for German clinics and medical practices. However, about 10 percent of these **masks** are unusable for medical purposes.
The big **refugee-mistake**: no labor market miracle has been brought by refugees. Unfortunately, most of the newcomers were not Syrian doctors and engineers.	The big **mistake about refugees**: no labor market miracle has been brought by refugees. Unfortunately, most of the newcomers were not Syrian doctors and engineers.	The big **mistake**: no labor market miracle has been brought by refugees. Unfortunately, most of the newcomers were not Syrian doctors and engineers.

Table 2: Example stimuli (translated into English from German). The variation between different conditions are marked in bold.

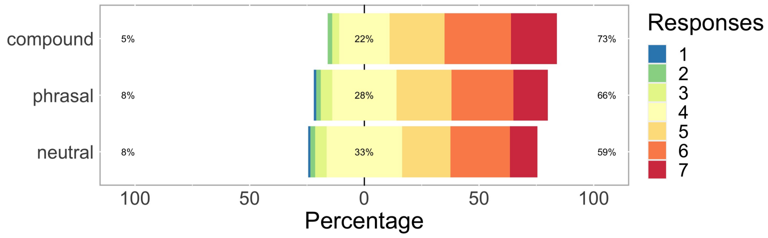


Figure 2: Distribution of participants' ratings by condition.

tasks (e.g., Brown et al., 2020; Chowdhery et al., 2023; Achiam et al., 2023; Touvron et al., 2023). However, the challenges associated with compound detection, particularly in identifying the associated attitudinal meanings of certain types of compounds like ECs remain significant. An avenue worth exploring is whether current LLMs encounter comparable challenges in this domain, particularly within the context of our psycholinguistic experiment. Recent work similar in spirit focused on human-likeness of LLMs' linguistic performance, e.g., testing language models on different syntactic phenomena,(Wilcox et al., 2018, 2020; Futrell et al., 2019; Arehalli et al., 2022) semantic judgements (e.g., Levy et al., 2017; Kauf et al., 2023), and on subtle pragmatic phenomena like irony or compliance with Gricean maxims (Hu et al., 2023; Tsvilodub et al., 2023).

We conducted experiments testing two of the latest versions of ChatGPT, namely GPT-4 and GPT-3.5-turbo (Achiam et al., 2023), employing various temperature settings. We designed a prompt that closely simulates the task employed in the experiment, and fed experimental items from the previous psycholinguistic experiment with human participants to these LLMs. Among these configurations, the one utilizing GPT-4 with a temperature set to 0 yielded the best results. Overall, we found that the best LLM captured a significant portion of the observed by-item variance in our experimental results ($R^2 = .48$, $p < .001$; see Fig. 3). Contrary to our experimental results, however, at the condition level, there was no indication of any alignment with human data ($R^2 = .43$, $p = .55$).

Our current LLM simulations thus provide initial evidence that these models currently have difficulty picking up cues for AMs conveyed by ECs. While further analyses (e.g., of the involved contextual embeddings or attention patterns) or future LLMs may provide a closer match between human ratings and modeling results, the current lack of effect was observed concurrently with the models' ability of capture substantial variation in other dimensions of our experimental results. This further highlights the specific subtleties and challenges involved in the detection and interpretation of ECs.

5. Recommendations and Outlook

We have now established that ECs systematically convey attitudinal meaning which can provide information for the NLP task of stance detection. We have also established that the current state of the art with respect to dependency parsers and UD treebank representations does not facilitate the automatic detection and identification of ECs. We furthermore showed that LLMs also struggle with the identification of EC contributions that are natural for humans, despite their otherwise impressive capabilities.

In this section, we propose that the UD community adopt a uniform approach towards the annotation of compounds. A systematic and uniform approach towards annotation will be able to result

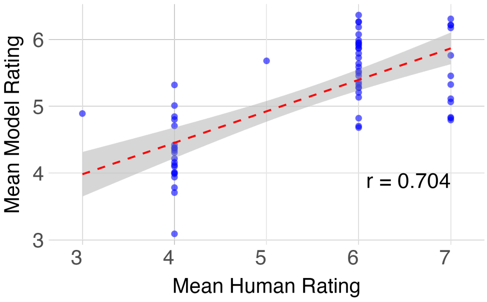


Figure 3: By-item correlation between participants' ratings from our experiment and LLM simulations.

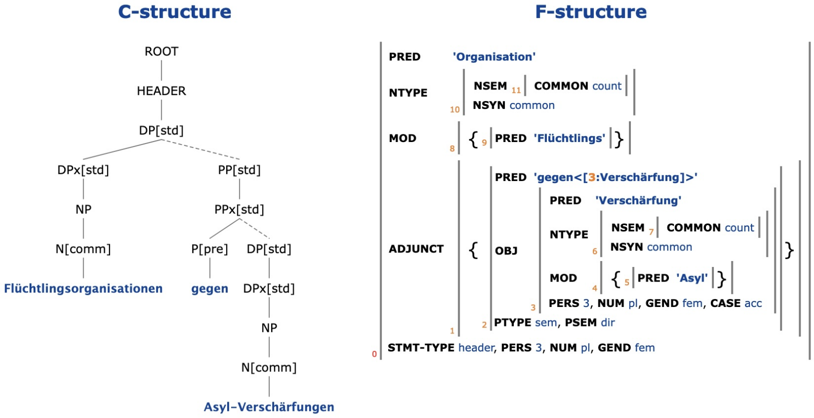


Figure 4: LFG analysis of ad hoc compounds.

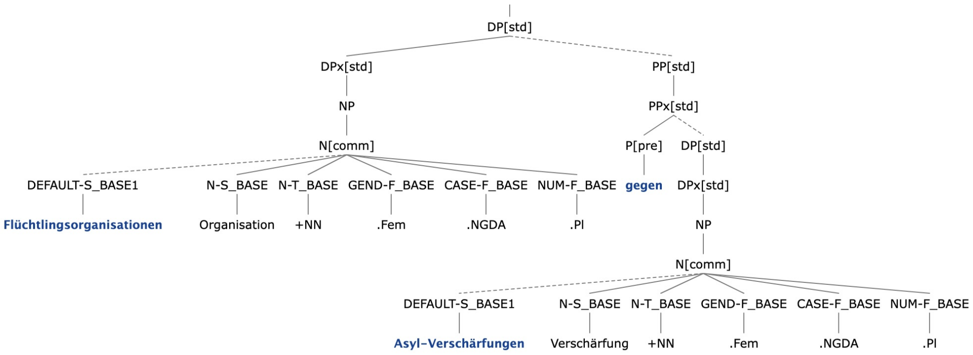


Figure 5: LFG analysis of ad hoc compounds with morphological analysis by DMOR.

in better down-stream machine learning and thus better results with respect to dependency parsers. Concretely we recommend adopting the approach deployed within the multilingual ParGram grammar development effort (Butt et al., 1999; Sulger et al., 2013). This is illustrated in Figures 4 and 5 from the German ParGram grammar (Dipper, 2003). The grammar is hosted on the INESS XLE website and can be used interactively.[10] The German ParGram grammar is based on Lexical Functional Grammar (LFG; Dalrymple, 2001), which has a context-free phrase structure part (the *c-structure*) and a dependency part (the *f-structure*). A c-structure of the compounds in question are simply tagged as common nouns (`N[comm]`). However, as shown in Figure 5, the German grammar also contains a finite-state morphological analyzer (DMOR, a precursor of SMOR; Schiller, 1994) and if one uses the built-in facility to look into the morphological analysis, one can see that the morphological analyzer separates out the parts of the compound into a base noun (the head noun) and the modifier, with the modifier then being flagged as such in the

[10] https://xle.uni-konstanz.de/iness/xle-web

dependency analysis at f-structure (Figure 4). We propose a UD annotation of the following form: a separation out of the head noun from the modifier, with the modifier being identified clearly as such in the dependency analysis. The curly brackets in the f-structure denote a set. This indicates that this attribute may have more than one value. Translating this into UD, we would assume that a head noun can have more than one modifier, all of which would be represented as sisters (at the same level) in the dependency graph.

However, a systematic annotation scheme only provides us with part of the necessary information for the detection of ad hoc compounds. Another part will necessarily involve the consultation of existing dictionaries, as was done as part of our corpus study (section 3.1). This type of lexical information can be further supplemented by lists of nouns and likely combinations, as was done in Schulte im Walde and Borgwaldt (2015). We propose that the data set we gleaned from the German newspaper study could be used in this way: one can compile an initial list of compounds for any given domain, identify the parts (i.e., heads and modifiers) of the compounds, and use the combined list of heads and modifiers as a seed list. This seed list can be then fed into models calculating clusters of lexically similar words for the identification of further ad hoc compounds. We leave this approach for exploration in further research.

6. Conclusion

We have presented a study of German ad hoc compounds that establishes that a subset of these compounds, dubbed *enigmatic compounds*, is systematically used to convey extra attitudinal meaning. We showed this via a combination of theoretical linguistic analysis, a corpus study and a psycholinguistic experiment. We also showed that the extra attitudinal meaning was predominantly used to express a negative stance in the newspapers and thus see enigmatic compounds as providing an important source of information for the end user NLP task of stance detection. A survey of existing dependency parsers and treebanks for German showed an uneven treatment for the annotation of German compounds and we therefore proposed a systematic annotation scheme that is based on the existing multilingual ParGram grammar development experience. We believe that a systematic annotation combined with lexical resources of the type developed in this paper will help ameliorate the challenge of automatized compound detection.

7. Acknowledgements

This project was funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under Germany's Excellence Strategy – EXC-2035/1 – 390681379.

8. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35:5113–5144.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Timothy Baldwin, William Croft, Joakim Nivre, Agata Savary, Sara Stymne, and Ekaterina Vylomova. 2023. Universals of linguistic idiosyncrasy in multilingual computational linguistics. *Dagstuhl Reports*, 13(4):22–70. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing (2nd ed.)*. Chapman and Hall/CRC.

David Beaver and Jason Stanley. 2018. Toward a non-ideal philosophy of language. *Graduate Faculty Philosophy Journal*, 39(2):503–547.

Reka Benczes. 2009. What motivates the production and use of metaphorical and metonymical compounds. *Cognitive approaches to English: Fundamental, methodological, interdisciplinary and applied aspects*, pages 49–69.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press, Cambridge. 2nd edition.

Douglas Biber and Edward Finegan. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, 9(1):93–124.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China. Coling 2010 Organizing Committee.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran & Associates, Inc.

Miriam Butt, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond. 1999. *A Grammar Writer's Cookbook*. CSLI Publications, Stanford.

Dennis Chong and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science*, 10:103–126.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24:1–113.

Rune Haubo B. Christensen. 2018. Cumulative link models for ordinal regression with the R package *ordinal*. *Journal of Statistical Software*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Mary Dalrymple. 2001. Lexical Functional Grammar. *Syntax and Semantics*, 34.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive-language. In *Proceedings of the 11th International Conference on Web and Social Media*, pages 512–515.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK. Coling 2008 Organizing Committee.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Stefanie Dipper. 2003. Implementing and documenting large-scale grammars – German LFG. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)*, 9(1).

John W. Du Bois. 2007. The stance triangle. In Robert Englebretson, editor, *Stancetaking in Discourse: Subjectivity, evaluation, interaction*, pages 139–182. John Benjamins, Amsterdam.

Mennatallah El-Assady, Valentin Gold, Annette Hautli-Janisz, Wolfgang Jentner, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2016. VisArgue - a visual text analytics framework for the study of deliberative communication. In *Proceedings of The International Conference on the Advances in Computational Analysis of Political Text (PolText2016)*, pages 31–36.

Mennatallah El-Assady, Wolfgang Jentner, Fabian Sperrle, Rita Sevastjanova, Annette Hautli-Janisz, Miriam Butt, and Daniel Keim. 2019. lingvis.io - a linguistic visual analytics framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Florence, Italy. Association for Computational Linguistics.

Gisbert Fanselow. 1981. Zur Syntax und Semantik der Nominalkomposition: ein Versuch praktischer Anwendung der Montague-Grammatik auf die Wortbildung im Deutschen. *Linguistische Arbeiten*, 107.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Esther Greussing and Hajo G. Boomgaarden. 2017. Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *Journal of ethnic and migration studies*, 43(11):1749–1774.

Robert Henderson and Elin McCready. 2019. Dog-whistles and the at-issue/non-at-issue distinction. In *Secondary content*, pages 222–245. Brill.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Alexandria J. Innes. 2010. When the threatened become the threat: The construction of asylum seekers in British media narratives. *International Relations*, 24(4):456–477.

IVW. 2023. Ranking der auflagenstärksten überregionalen Tageszeitungen in Deutschland im 2. Quartal 2023.

Carina Kauf, Anna A. Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan S. She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11).

Christopher Kennedy. 1999. *Projecting the adjective: the syntax and semantics of gradability and comparison*. Garland.

Hans J. Kleinsteuber and Barbara Thomass. 2007. The German media landscape. *European Media Governance: National and regional dimensions*, pages 111–123.

Judith N. Levi. 1978. *The syntax and semantics of complex nominals*. Academic Press, New York.

Joseph Patrick Levy, John Bullinaria, and Samantha McCormick. 2017. Semantic vector evaluation and human performance on a new vocabulary MCQ test. In *Proceedings of the Annual Conference of the Cognitive Science Society: CogSci 2017 London: "Computational Foundations of Cognition"*. Cognitive Science Society.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6299–6305, Hong Kong, China. Association for Computational Linguistics.

Ralf Meyer. 1993. Compound comprehension in isolation and in context. In *Compound Comprehension in Isolation and in Context*. Max Niemeyer Verlag, Tübingen.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Gregory Murphy. 2002. *The Big Book of Concepts*. MIT Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Mary Ellen Ryder. 1994. *Ordered chaos: The interpretation of English noun-noun compounds*, volume 123. University of California Press, Berkeley.

Galit Weidmann Sassoon. 2011. Adjectival versus nominal categorization processes: The rule vs. similarity hypothesis. *Belgian Journal of Linguistics*, 25:104–147.

Anne Schiller. 1994. DMOR - User's Guide. Technical report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *Künstliche Intelligenz*, 35:329–341.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sabine Schulte im Walde and Susanne Borgwaldt. 2015. Association norms for German noun compounds and their constituents. *Behavior Research Methods*, 47(4):1199–1221.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A new hybrid dependency parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology*, pages 115–124.

Sebastian Sulger, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh M. Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Agnieszka Patejuk, Özlem Çetinoglu, I Wayan Arka, and Meladel Mistica. 2013. ParGramBank: The ParGram parallel treebank. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 550–560, Sofia. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Polina Tsvilodub, Michael Franke, Robert Hawkins, and Noah D. Goodman. 2023. Overinformative question answering by humans and machines. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, page 1707–1713.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Wolfgang Wildgen. 1981. *Makroprozesse bei der Verwendung nominaler ad hoc-Komposita im Deutschen*. Linguistic Agency University of Trier.

Malte Zimmermann. 2011. Discourse particles. In Paul Portner, Claudia Maienborn, and Klaus von Heusinger, editors, *Semantics: An International Handbook of Natural Language Meaning*, pages 2011–2038. De Gruyter Mouton, Berlin.

Author Index