COLING 2025

18th Workshop on
Building and Using Comparable Corpora

PROCEEDINGS

20 January, 2025

©Copyright The International Committee on Computational Linguistics (ICCL), 2025
These proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 979-8-89176-211-4
ISSN 2951-2093 (COLING)

Message from the Program Chairs

The 18th Workshop on Building and Using Comparable Corpora
(BUCC) @ COLING 2025

In the language engineering and linguistics communities, research in comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is chiefly motivated by the need to use comparable corpora as training data for statistical NLP applications such as statistical and neural machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest because they enable cross-language discoveries and comparisons. It is generally accepted in both communities that comparable corpora consist of documents that are comparable in content and form in various degrees and dimensions across several languages or language varieties. Parallel corpora are on the one end of this spectrum, unrelated corpora on the other.

Comparable corpora have been used in various applications, including Information Retrieval, Machine Translation, Cross-lingual text classification, etc. The linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for statistical natural language processing applications, for example, to extract parallel corpora from comparable corpora for neural machine translation. As such, it is of great interest to bring together builders and users of such corpora. The aim of the workshop series on "Building and Using Comparable Corpora" (BUCC) is to promote progress in this field.

The previous editions of the workshop took place in Africa (LREC 2008 in Marrakech), America (ACL 2011 in Portland and ACL 2017 in Vancouver), Asia (ACL-IJCNLP 2009 in Singapore, ACL-IJCNLP 2015 in Beijing, LREC 2018 in Miyazaki, Japan), Europe (LREC 2010 in Malta, ACL 2013 in Sofia, LREC 2014 in Reykjavik, LREC 2016 in Portoroz, RANLP 2019 and RANLP 2023 in Varna, LREC 2022 in Marseille, LREC-COLING-2024 in Turin) and also on the border between Asia and Europe (LREC 2012 in Istanbul). Due to the Corona crisis, the workshop was also held online in conjunction with LREC 2020 and RANLP 2021. The materials of the past workshops and related studies have also been summarised in a recent textbook from Springer:
`https://link.springer.com/book/10.1007/978-3-031-31384-4`.

We want to thank all the people who, in one way or another, helped make this workshop once again a success, especially the COLING workshop chairs, and publication chairs.

Our special thanks go to our invited speakers, Ken Church and Preslav Nakov, and to the members of the program committee, who did a great job in reviewing the submitted papers under strict time constraints. Last but not least, we would like to thank the authors and all workshop participants.

Serge Sharoff, Ayla Rigouts Terryn, Pierre Zweigenbaum, Reinhard Rapp   January 2025

Organizing Committee

Serge Sharoff University of Leeds, United Kingdom

Ayla Rigouts Terryn Université de Montréal, Mila, Canada

Pierre Zweigenbaum Université Paris-Saclay, CNRS, LISN, Orsay, France

Reinhard Rapp University of Mainz, Germany

Program Committee

- Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
- Eleftherios Avramidis (DFKI, Germany)
- Gabriel Bernier-Colborne (National Research Council, Canada)
- Thierry Etchegoyhen (Vicomtech, Spain)
- Natalia Grabar (University of Lille, France)
- Amal Haddad Haddad (Universidad de Granada, Spain)
- Kyo Kageura (University of Tokyo, Japan)
- Natalie Kübler (Université Paris Cité, France)
- Philippe Langlais (Université de Montréal, Canada)
- Yves Lepage (Waseda University, Japan)
- Michael Mohler (Language Computer Corporation, USA)
- Emmanuel Morin (Nantes Université, France)
- Dragos Stefan Munteanu (Language Weaver, Inc., USA)
- Ted Pedersen (University of Minnesota, Duluth, USA)
- Nasredine Semmar (CEA LIST, Paris, France)
- Silvia Severini (Leonardo Labs, Italy)
- Richard Sproat (OGI School of Science & Technology, USA)
- Tim Van de Cruys (KU Leuven, Belgium)
- François Yvon (Sorbonne Université, France)

Table of Contents

Conference Program

9:15–9:30 **Opening and introduction**

9:30–10:30 **Multilingual corpus development**

Bilingual resources for Moroccan Sign Language Generation and Standard Arabic Skills Improvement of Deaf Children
Abdelhadi Soudi, Corinne Vinopol and Kristof Van Laerhoven

Harmonizing Annotation of Turkic Postverbial Constructions: A Comparative Study of UD Treebanks
Arofat Akhundjanova

10:30–11:00 **Coffee break, morning**

11:00–13:00 **Multilinguality of Large Language Models**

Towards Truly Open, Language-Specific, Safe, Factual, and Specialized Large Language Models
Preslav Nakov

Make Satire Boring Again: Reducing Stylistic Bias of Satirical Corpus by Utilizing Generative LLMs
Asli Umay Ozturk, Recep Firat Cekinel and Pinar Karagoz

BEIR-NL: Zero-shot Information Retrieval Benchmark for the Dutch Language
Ehsan Lotfi, Nikolay Banar and Walter Daelemans

13:00–14:00 Lunch

14:00–15:30 Machine Translation and Cross-lingual Processing

Refining Dimensions for Improving Clustering-based Cross-lingual Topic Models
Chia-Hsuan Chang, Tien Yuan Huang, Yi-Hang Tsai, Chia-Ming Chang and San-Yih Hwang

The Role of Handling Attributive Nouns in Improving Chinese-To-English Machine Translation
Adam Meyers, Rodolfo Joel Zevallos, John E. Ortega and Lisa Wang

Can a Neural Model Guide Fieldwork? A Case Study on Morphological Data Collection
Aso Mahmudi, Borja Herce, Demian Inostroza Améstica, Andreas Scherbakov, Eduard H. Hovy and Ekaterina Vylomova

15:30–16:00 Coffee break, afternoon

16:00–17:30 Diversity of language resources

Comparable Corpora: Opportunities for New Research Directions
Kenneth Ward Church

SELEXINI – a large and diverse automatically parsed corpus of French
Manon Scholivet, Agata Savary, Louis Estève, Marie Candito and Carlos Ramisch

17:30–17:45 Closing remarks

Bilingual resources for Moroccan Sign Language Generation and Standard Arabic Skills Improvement of Deaf Children

Abdelhadi Soudi
Ecole Nationale Supérieure
des Mines de
Rabat, Morocco
asoudi@enim.ac.ma

Corinne Vinopol
Institute for Disabilities
Research and Training
United States
corinne@idrt.com

Kristof Van Laerhoven
University of Siegen, Germany
kvl@eti.uni-siegen.de

Abstract

This paper presents a set of bilingual Standard Arabic (SA)-Moroccan Sign Language (MSL) tools and resources to improve Moroccan Deaf children's SA skills. An MSL Generator based on rule-based machine translation (MT) is described that enables users and educators of Deaf children, in particular, to enter Arabic text and generate its corresponding MSL translation in both graphic and video format. The generated graphics can be printed and imported into an Arabic reading passage. We have also developed MSL Clip and Create software that includes a bilingual database of 3,000 MSL signs and SA words, a Publisher for the incorporation of MSL graphic support into SA reading passages, and six Templates that create customized bilingual crossword puzzles, word searches, Bingo cards, matching games, flashcards, and fingerspelling scrambles. A crowdsourcing platform for MSL data collection is also described. A major social benefit of the development of these resources is in relation to equity and the status of deaf people in Moroccan society. More appropriate resources for the bilingual education of Deaf children (in MSL and SA) will lead to improved quality of educational services.

1 Introduction

Research into brain processing of language has shown that signed and spoken languages occur in the same region of the brain, but there are differences according to language modality of representation (Campbell et al., 2007). This information has implications for designing Natural Language Processing (NLP) systems to facilitate Deaf individuals' access to education. Consideration must be given to language input and output modalities and representations.

Spoken languages rely on audition. Simplistically stated, to express Standard Arabic (SA) in written form, we choose letters that are graphic representations of sounds and put them in that oral production order. When we choose to record or retrieve a word, we can use its spelling and alphabetic sequence.

On the other hand, if a person must rely on vision to develop language, visual principles affect how that language is organized and expressed. Expression tends to be primarily manual and facial, and sign languages (SLs) incorporate many techniques that visually and kinesthetically convey life's experiences. As a result, to express a sign language, it must be depicted through graphics, animation, or video.

Stokoe (1960) coined the terms, chereme and cherology, from the Greek word χεíρ for hand. He considered a chereme a basic unit of signed communication, functionally and psychologically equivalent to the phonemes of oral languages. He posited that signs can be described primarily by four cheremes, classified as tab (elements of location from the Latin tabula), dez (the hand shape, from designator), sig (the motion, from signation), and with some researchers, ori (orientation). Facial expression and mouthing are also phonemic in sign language. There have been a few attempts to develop a written form to describe sign language (e.g., SignWriting), but these are hardly used or recognized by Deaf people or their service providers.

2 Deafness and Moroccan Sign Language (MSL)

According to Morocco's High Commission of Planning's 2014 Census (El Ouazzani, 2015) and a survey on disability statistics conducted by the Ministry of Solidarity, Women, Family and Social Development in April 2014 (Lkhoulf, 2017) 3.5% of the population (1,182,681 people) have some degree of hearing loss, 0.2% (56,745 people) have a profound hearing loss or total inability to hear, 1.0% (347,386 people) have "a lot of difficulty"

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 1–9
January 20, 2025. ©2025 Association for Computational Linguistics

hearing, and 2.3% (778,550 people) have “some” difficulty hearing. There is, therefore, need for both SL use and Deaf education. However, education of Deaf children in Morocco is very dire. Approximately 85% do not attend school. Education beyond sixth grade until recently was unavailable, and very few are gainfully employed. The plight of educating Deaf children is further compounded by other issues (Soudi and Vinopol, 2019).

First, the language of instruction is Arabic/French audio and text-based and without interpretation into MSL unless a volunteer “interpreter” from a Deaf Association is present. Most Deaf children use MSL which is an independent gestural system of communication that does not rely on audition but does, to a great extent, on the logic of the visual experience. It is not an interpretation of SA or spoken vernaculars. It can only be depicted through graphics, video, or animation. Numerous studies have demonstrated that teaching Deaf students is best achieved bilingually (i.e., through both their native signed and spoken languages).

Second, there is a lack of well-trained educators of Deaf children who are familiar with the metacognitive skills essential for effective reading comprehension. Those who can communicate with the children in SL have little training or understanding about how to make educational content meaningful to them.

Third, almost no sign language interpreters exist to help include Deaf children in the regular curriculum with hearing peers. Deaf children who do attend school are kept in segregated classrooms. Therefore, there is little opportunity for Deaf children to get the breadth of educational information that their hearing peers have.

3 Bilingual resources for MSL Generation and SA Skills Vocabulary Building and Improved Reading Comprehension

In view of the challenges outlined above, the development of tools and resources to help Moroccan Deaf children improve their SA skills and access to education is badly needed. In this section, we describe two software programs that we have developed, namely (i) an MSL generator based on rule-based machine translation and (ii) MSL Clip and Create, a set of tools and bilingual resources for custom creation of MSL-supported instructional materials for the improvement of SA skills of Moroccan Deaf children.

3.1 MSL Generator

Several studies have demonstrated that a combination of signed spoken pictures/graphics and comprehension of written text can facilitate Deaf students’ spoken language skills and provide support for word recognition (Nielsen et al., 2016; Wilson and Hyde, 1997). Wilson and Hyde (1997) reported that the use of Signed English reading books significantly improves reading comprehension of Deaf students. Similarly, Nielsen et al. (2016) and Stryker et al. (2015) argue that the use of Signed Exact English (SEE) supports the comprehension of reading by Deaf children.

In this subsection, we describe an MSL Generator that uses rule-based machine translation (MT). The system generates sign graphic and video supports for SA text. Teachers can print the graphics and incorporate them into reading passages. Educators of Deaf children can also combine the use of the Generator with the use of the MSL Clip and Create software as described below. That software has sign concepts/images that they can import from the database and add to the reading text.

Research on sign language machine translation (SLMT) is novel compared to research on spoken language MT. Work on SLMT faces some problems, including the lack of parallel corpora, a formal writing system of SLs, and a standard representation format.

Examples of rule-based SLMT systems include the Zardoz system, which translates English text into Japanese Sign Language and ASL (Veale et al., 1998); the Albuquerque Weather System, which translates from English text to ASL in the weather forecast domain (Grieve-Smith, 1999): the TEAM Project (Zhao et al., 2000), which translates English into ASL; and the Greek-to-Greek Sign Language System (Kouremenos et al., 2018).

Data-driven approaches to SLMT include, to name a few, Ebling (2016)’s automatic translation from German to Synthesized Swiss German Sign Language and Bauer et al. (1999)’s statistical-based SL translation system, which translates from recognized video-based continuous SL (German Sign Language) to spoken language (German) in the domain of shopping.

While current SLMT research tends to use data-driven approaches, most (if not all) existing systems either translate in a limited domain or are not actually used by the Deaf Community in real-life situations. This is largely because data-driven ap-

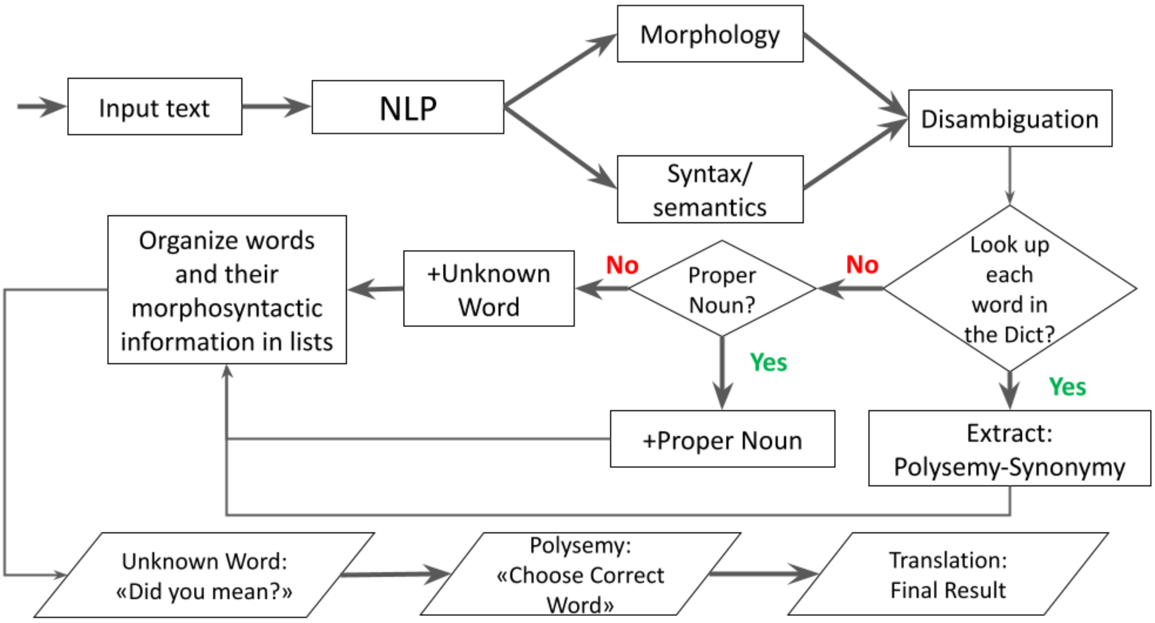

Input text
NLP
Morphology
Syntax/ semantics
Disambiguation
Look up each word in the Dict?
No
Proper Noun?
No
+Unknown Word
Organize words and their morphosyntactic information in lists
Yes
+Proper Noun
Yes
Extract: Polysemy-Synonymy
Unknown Word: «Did you mean?»
Polysemy: «Choose Correct Word»
Translation: Final Result

Figure 1: MSL Generator Process Logic

proaches require large-scale parallel corpora that, unfortunately, are not as available for sign languages as they are for spoken languages (Jiang et al., 2023). In addition, most systems do not accommodate SL regional variations.

Using insight gained from previous data-driven SLMT research experiences, we chose to build a rule-based MT (RBMT) system. With this approach, we were able to handle SL regional variations and word sense disambiguation and enrich the current MSL database. Another major reason for choosing a rule-based approach is that it is inclusive of minority or low-resourced languages, such as MSL. Our approach agreed with Hurskainen and Tiedemann (2018, p. 1) who argue that “if we use statistical or neural translation systems, we will exclude 99.8 percent of languages out of development.” They further state that "the current hype on neural methods still accelerates the break between the small group of dominant languages and the less-resourced ones. If we want to avoid the break, we do not see any other way out than to put efforts in developing such systems that are affordable for less-resourced languages.”

RBMT systems generally handle translation by parsing the source text. The analysis module then results in an intermediate representation that acts as input to the generation module that generates the translation in the target language. Mappings between lexical items stored in bilingual dictionaries and transfer rules are applied to account for the linguistic mismatches between the source language and target language.

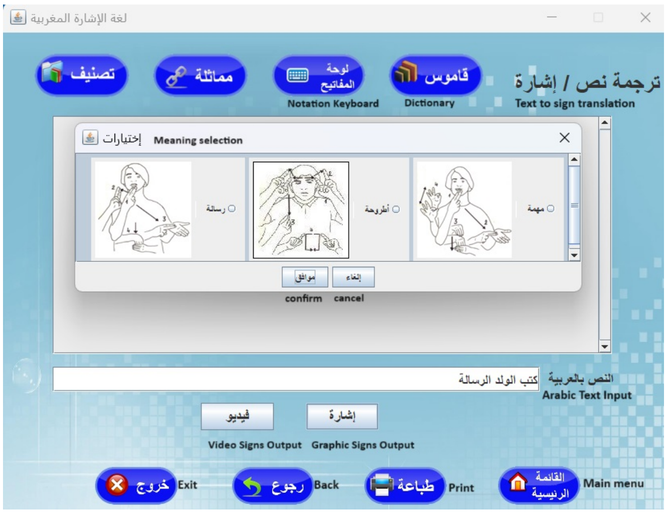
Figure 1 shows the MSL Generator logic process. The input to the MSL Generator is Arabic text that is tokenized and then processed using:

(i) Buckwalter Arabic morphological Analyzer (BAMA) (Buckwalter, 2002). BAMA has a lexicon of 40,648 lemmas and three morphological compatibility tables used for controlling affix-stem combinations.

(ii) Stanford parser[1], a statistical parser with pretrained models for English, German, and Arabic.

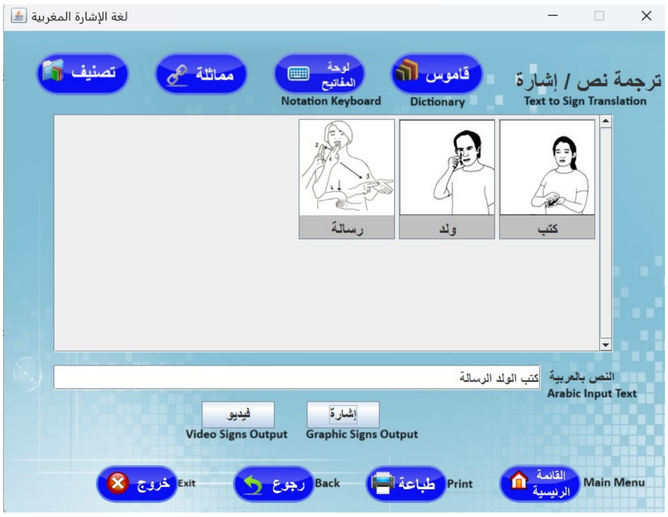
The MSL Generator then looks up every word in the database. If the word exists in the dictionary, its corresponding sign video and/or graphic (depending on the user’s preference) are retrieved. If the word is polysemous, its distinct meanings are also retrieved and displayed on the output screen, each with its corresponding graphic sign. The user then chooses the correct meaning. If the word is not found in the database, the system checks if it can be found in the Named Entity dictionary. If it is not found, the system returns “unknown word” or suggests a possible word(s) based on the context by asking the user “Did you

[1] http://nlp.stanford.edu/software/lex-parser.shtml



كتب الولد الرسالة

Figure 2: Screenshot of the generation of a sentence that includes a polysemous word



كتب الولد الرسالة

Figure 3: Signed Standard Arabic generation after selection of the correct meaning

mean. . . . . .”. If all words are found, the user can display the translation of the Arabic text either in the form of graphic sign or video sign.

Figure 2 illustrates the translation of the Arabic sentence كتب الولد الرسالة “The boy wrote the letter.” In this sentence, the word “الرسالة” is polysemous. It has three distinct meanings “dissertation,” “letter,” and “mission.” Accordingly, as can be seen in Figure 2, the system displays the three distinct meanings, each with its corresponding graphic sign, and asks the user to choose the intended meaning.

After selection of the correct and intended meaning of the word الرسالة, the system generates the corresponding translation of the input text in MSL, as is shown in Figure 3. Users can also print the sign graphics output.

We are currently developing fragments of MSL grammar through a large MSL corpus that is being created and annotated. We will integrate fragments of MSL grammar into the Generator as much as possible.

We are also incorporating another output mode that will allow MSL Generator to instantly return a pre-made sign language video or avatar sign sequence matching the input text. For this purpose, we use an XML description language (SiGML), which is based on HamNoSys notation (Hanke and Schmaling, 2001), and Lebourque and Geibet’s gesture specification language (GessyCA) (Lebourque and Gibet, 1999). A system was developed to convert HamNoSys code of the given word to its SiGML form, to enable the avatar animation.

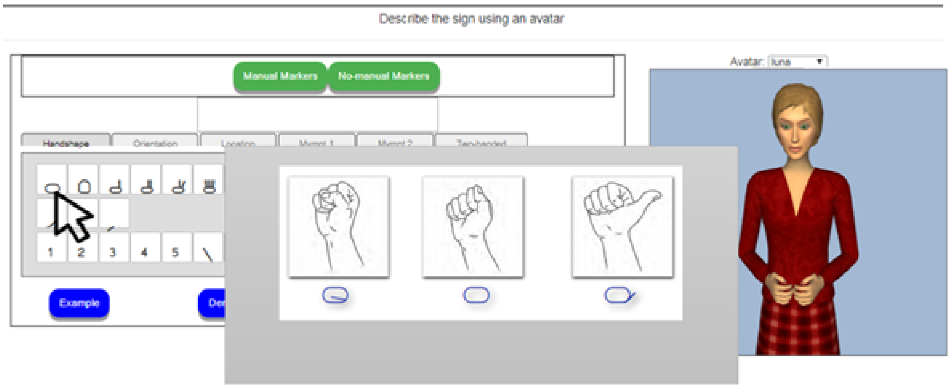


Figure 4: The enhanced version of HamNoSys for the description of signs to animate

To facilitate the animation process, we have created a notation keyboard based on an enhanced version of HamNosys that is legible and user-friendly. HamNosys symbols are enhanced with a pictorial list of the manual and non-manual markers as shown in Figure 4. We are in the process of animating 3,000 SA words in the dictionary used for the MSL Generator.

3.2 MSL Clip and Create Software

MSL Clip and Create is a set of apps, SL media assets, and bilingual resources for the customized creation of MSL-supported instructional material to improve Deaf learners's SA skills. The software includes the following components:

- A bilingual database of 3,000 SA words-MSL signs. Each sign has a corresponding definition in both MSL and SA. Two methods were used to collect data for the database:

  (i) Mobile recording studios: In the first phase, suitable participants from five regions of Morocco (i.e., Souss-Massa, Rabat-Salé-Kénitra, Tanger-Tétouan-Al Hoceima, Marrakech-Safi, and Fes-Meknès) are recruited based on demographic information collected through a detailed questionnaire. The latter includes, inter alia, information on the informants' sex, age, age of MSL acquisition, type of Deafness, frequency of use of MSL, and education. A committee, consisting of a focal team member and local coordinators belonging to a major Deaf association in each region, selects the candidates for the data lexical elicitation task. Following Schembri et al. (2012), Milroy (1980), and Bayley et al. (2001), we involve local Deaf signers from regional associations to both recruit local informants and lead the data collection process. This method ensures that the data collection tasks are Deaf-friendly and adapt to the sociocultural experiences of the participants. The importance of involving Deaf assistants in the process of data collection has been documented in the literature (Harris et al., 2009; Ladd, 2003; Singleton et al., 2012). These authors also recommend that the required informed consent be translated into the participants' native language. This recommendation is particularly necessary in Morocco, where the Deaf community has very limited spoken language proficiency. Using this method, we were able to initially collect 2,200 signs.

  (ii) In order to accommodate regional sign variations across Morocco and to meet the needs of educators of Deaf children, we developed a crowd-sourcing platform, Madrasati-Signs «My school signs» [2] that includes a database of 3,000 words categorized into 21 domains (i.e, family, colors, home, food, clothes, time, sports and hobbies, body, feelings, geography, transportation, education, nature and weather, animals, health, math, business and careers, media and arts, technology, alphabet, life sciences and physics). The choice of words is based on an enhanced version of MacArthur-Bates Communicative Development Inventory Words and Moroccan (STEM)

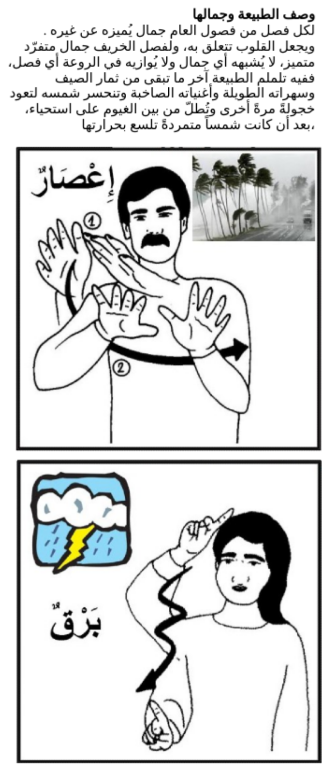
[2] https://madrasati-signs.org/

textbooks (Fenson et al., 2007). Most words in the database have a corresponding concept. The participants were asked to record signs for words (and concepts) in a specific domain using the Madrasati-Signs platform. The platform is designed in a similar way as the crowdsourcing platform AfricaSign (Soudi et al., 2019). Madrasati-Signs can accommodate the recording of multiple signs (if any) for each word[3]. Where regional variations exist for a single sign, they are identified as (sign 1), (sign 2), and so forth. Users have two input modes:

a. Add their signs by videotaping them using Laptop/Phone cam. After their consent, the users' phone/laptop cams will be automatically activated, and they will be asked to provide a sign for a particular word. They will then have the possibility of viewing their sign, and either validate it or videotape it again. Figure 6 shows a Deaf signer from Rabat recording signs in the Home domain.

b. Uploading a video sign if a user already has it.

For quality assurance purposes, sign contributions to the platform were restricted to trusted signers selected by regional Deaf associations. Madrasati-signs users logged in and described themselves demographically by region and Deafness affiliation (e.g., Deaf themselves, have Deaf parents-CODA).



وصف الطبيعة وجمالها
لكل فصل من فصول العام جمال يُميزه عن غيره .
ويجعل القلوب تتعلق به، ولفصل الخريف جمال متفرّد
متميز، لا يُشبهه أي جمال ولا يُوازيه في الروعة أي فصل،
ففيه تلملم الطبيعة آخر ما تبقى من ثمار الصيف
وسهراته الطويلة وأغنياته الصاخبة وتنحسر شمسه لتعود
خجولةً مرةً أخرى وتُطلّ من بين الغيوم على استحياء،
،بعد أن كانت شمساً متمردةً تلسع بحرارتها

Figure 5: A screenshot of the Publisher illustrating an imported reading passage on Nature with graphic sign supports from the database

- MSL Clip and Create also includes a publishing tool for the creation of customized and printable materials using the graphics in the database. Users can also import other graphics and photos from their device. Educators of Deaf children, for example, can import a SA reading passage from a Moroccan textbook and support it with graphic signs and concepts from the database, as is shown in Figure 5. The user can import as many sign graphics as they desire in real time. This tool is particularly useful in education environments in which textbooks are designed for the hearing and are not adapted to the needs of Deaf learners, as in the case of Morocco.

- In addition to the database, and publisher, the software also has six templates that instantly create bilingual SA-MSL customized crossword puzzles, word searches, bingo cards, matching games, flashcards, and fingerspelling scrambles using any graphic signs of the database. Figure 8 shows a screenshot of a customized bingo card and fingerspelling scramble puzzle.

- The software also includes a story-builder that currently hosts three Hispanic folktales translated into SA and three stories from a Moroccan national SA textbook. Users can view the story in MSL and/or read the Arabic text. The latter can be automatically diacritized with a simple click of a button. This functionality is necessary for early grade children who still do not know the grammar of SA and, therefore, cannot read it without diacritics. Figure 7 shows a screenshot of one of the stories in the MSL Clip and Create software.

These tools and resources can be used for

[3] For the list of words and their corresponding concepts that were included in the lexical elicitation task, see www.madrasati-signs.org

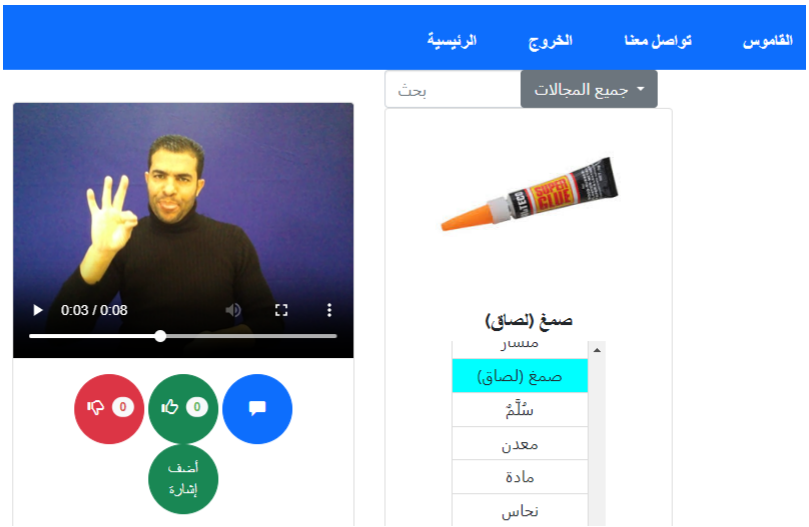
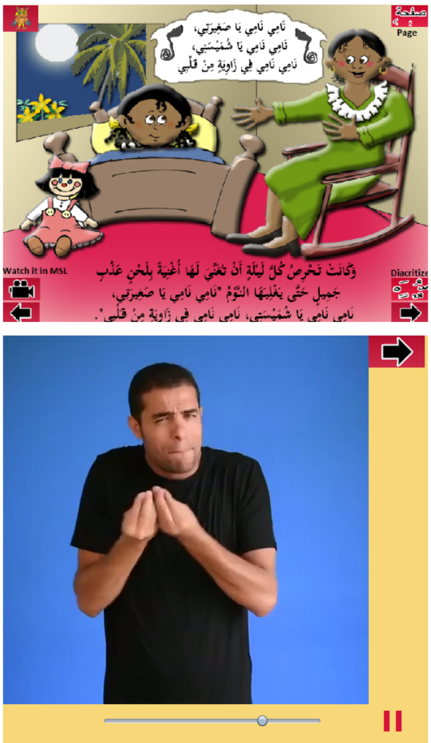


Figure 6: A screenshot of a Deaf signer from Rabat recording signs in the Home domain using Madrasati-Signs platform


صفحة
2
Page
نَامِي نَامِي يَا صَغِيرَتِي،
نَامِي نَامِي يَا شُمَيْسَتِي،
نَامِي نَامِي فِي زَاوِيَةٍ مِنْ قَلْبِي
Watch it in MSL
Diacritize

وَكَانَتْ تَحْرِصُ كُلَّ لَيْلَةٍ أَنْ تُغَنِّيَ لَهَا أُغْنِيَةً بِلَحْنٍ عَذْبٍ
جَمِيلٍ حَتَّى يَغْلِبَهَا النَّوْمُ "نَامِي نَامِي يَا صَغِيرَتِي،
نَامِي نَامِي يَا شُمَيْسَتِي، نَامِي نَامِي فِي زَاوِيَةٍ مِنْ قَلْبِي".

Figure 7: A screenshot of one of the Software's stories

reading-related activities, such as SA vocabulary building and word recognition in each of the 21 dictionary domains.

4 Broader Impact

In addition to academic benefits, a major social benefit of the development of these MSL-SA resources is in relation to equity and the status of Deaf people in Moroccan society. More appropriate resources for the bilingual education of Deaf children (in MSL and Arabic) will lead to improved quality of educational and interpreting services for Deaf people and provide more opportunities for self-development and employment. Deaf people who can become more highly qualified and trained will be in a better position to contribute to society in different ways, and will be able to achieve greater recognition, access, and equity in the wider community. Furthermore, the greater understanding of MSL and improved resources for MSL teaching, learning and research can provide an evidence-base for policy-makers in supporting appropriate education, training and services for Deaf children and adults. In this context, it is worth noting that the Moroccan Ministry of Education has endorsed our tools and resources and helped with their free distribution to Deaf associations across Morocco. The Ministry is currently investigating establishment of a teacher of the Deaf training program. These efforts will help close the gap in education, employment, and health between Deaf people throughout their lifespan and their hearing peers.

5 Limitations

As is the case of most other languages, one of the major limitations of the MSL Generator is the lack of a comprehensive grammar of MSL. SLs are natural languages, and they have also developed linguistic systems with a grammar and a vocabulary (Johnston and Schembri, 2007). However, there is no other SL that has a reference grammar "that meets

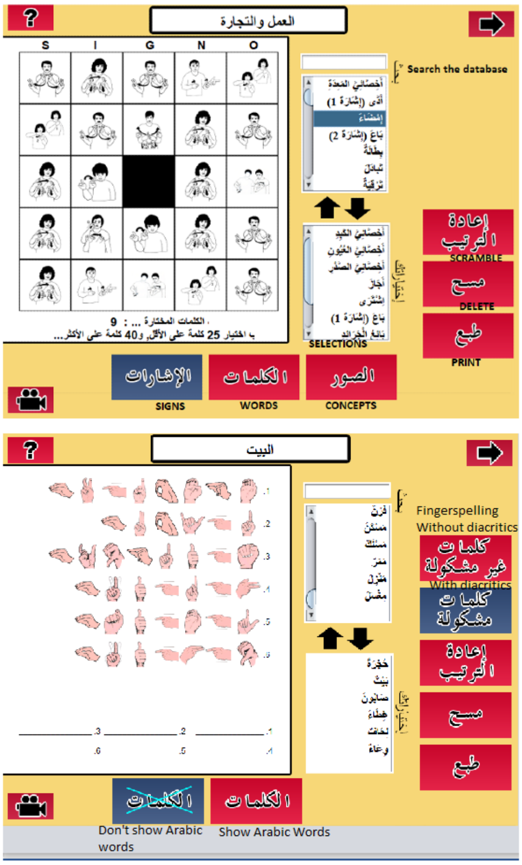


Figure 8: Screenshots of bingo and fingerspelling templates

the common standards set by spoken language reference grammars" (Palfreyman et al., 2015). We are currently addressing this limitation though the creation of a large scale MSL corpus which is an important resource to understand the grammar of MSL. Data is being collected from a total of 240 Deaf signers from the twelve regions of Morocco. The dataset, totaling 120 hours of video, is being/will be translated, of which 10 hours are being annotated, and tagged using EUDICO Linguistic Annotator (ELAN). Fragments of MSL grammar will be incorporated into MSL Generator. The corpus creation will also enrich the SA-MSL database and help understand the sociolinguistic situation of MSL by investigating factors, such as the multilingual linguistic environment, gender, regional variation, family and education.

6 Conclusion

This paper has presented a set of SA and MSL tools and resources calculated to improve Moroccan Deaf children's SA skills. We have described two self-developed resources that are intended to help Deaf learners improve their SA skills. MSL Generator enables educators of Deaf children to enter Arabic text and generate its corresponding translation in both MSL graphic and video formats. The generated graphics can be printed and imported into an Arabic reading passage. We have also described MSL Clip and Create software, which includes a database of 3,000 SA words and MSL signs, a Publisher for the incorporation of MSL support into Arabic reading passages, and six Templates that create customized crossword puzzles, word searches, Bingo cards, matching games, flashcards, and fingerspelling scrambles.

Helping Deaf children improve their SA skills is challenging and requires a strong long-term commitment, particularly in light of the lack of resources available in their native sign language.

References

Britta Bauer, Sonja Nießen, and Hermann Hienz. 1999. Towards an automatic sign language translation system. In *Proceedings of the international workshop on physicality and tangibility in interaction: towards new paradigms for interaction beyond the desktop, Siena, Italy*.

Robert Bayley, Clayton Valli, and Ceil Lucas. 2001. *Sociolinguistic variation in American sign language*. Gallaudet University Press.

Tim Buckwalter. 2002. *Buckwalter Arabic Morphological Analyzer Version 1.0*. Linguistic Data Consortium, University of Pennsylvania. LDC Catalog No.: LDC2002L49.

Ruth Campbell, Mairéad MacSweeney, and Dafydd Waters. 2007. Sign language and the brain: a review. *Journal of deaf studies and deaf education*, 13(1):3–20.

Sarah Ebling. 2016. *Automatic Translation from German to Synthesized Swiss German Sign Language*. Ph.D. thesis, University of Zurich.

Z. El Ouazzani. 2015. Moroccan experience on disability statistics. Washington Group Meeting, Copenhagen, Denmark. Retrieved from https://www.cdc.gov/nchs/data/washington_group/meeting15/wg15_session_8_4_touahami.pdf.

Larry Fenson et al. 2007. Macarthur-bates communicative development inventories.

Angus B Grieve-Smith. 1999. English to american sign language machine translation of weather reports. In *Proceedings of the Second High Desert Student Conference in Linguistics (HDSL2), Albuquerque, NM*, pages 23–30.

T Hanke and C Schmaling. 2001. A hamnosys-based phonetic transcription system as a basis for sign language generation. In *Gesture Workshop 2001*.

Raychelle Harris, Heidi M Holmes, and Donna M Mertens. 2009. Research ethics in sign language communities. *Sign Language Studies*, 9(2):104–131.

Arvi Hurskainen and Jörg Tiedemann. 2018. Rule-based machine translation from english to finnish. In *Proceedings of the Second Conference on Machine Translation (WMT2017)*. The Association for Computational Linguistics.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Trevor Johnston and Adam Schembri. 2007. *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge University Press.

Dimitrios Kouremenos, Klimis Ntalianis, and Stefanos Kollias. 2018. A novel rule based machine translation scheme from greek to greek sign language: Production of different types of large corpora and language models evaluation. *Computer Speech & Language*, 51:110–135.

Paddy Ladd. 2003. *Understanding deaf culture: In search of deafhood*. Multilingual Matters.

Thierry Lebourque and Sylvie Gibet. 1999. High-level specification and control of communication gestures: The gessyca system. In *Computer Animation, 1999. Proceedings*, pages 24–35.

N. Lkhoulf. 2017. Disability statistics from the 2014 moroccan census. In *Regional Meeting on Disability Measurement and Statistics in support of the 2030 Agenda for Sustainable Development and the 2020 World Population and Housing Census Programme*, Muscat, Oman.

Lesley Milroy. 1980. Language and social networks. *(No Title)*.

Diane Corcoran Nielsen, Barbara Luetke, Meigan McLean, and Deborah Stryker. 2016. The english-language and reading achievement of a cohort of deaf students speaking and signing standard english: A preliminary study. *American Annals of the Deaf*, 161(3):342–368.

Nick Palfreyman, Keiko Sagara, and Ulrike Zeshan. 2015. Methods in carrying out language typological research. In Eleni Orfanidou, Bencie Woll, and Gary Morgan, editors, *Research Methods in Sign Language Studies: A Practical Guide*, pages 173–192. Wiley-Blackwell, Chichester, UK.

Adam Schembri, Jordan Fenlon, Ramas Rentelis, Steve Reynolds, and Kearsy Cormier. 2012. Towards a british sign language corpus: A short report. *International Journal of Corpus Linguistics*, 17(1):3–15.

Jenny L Singleton, Gabrielle Jones, and Shilpa Hanumantha. 2012. Toward ethical research practice with deaf participants. *Journal of Empirical Research on Human Research Ethics*, 9(3):59–66.

Abdelhadi Soudi, Kristof Van Laerhoven, and Elmostafa Bou-Souf. 2019. Africasign – a crowd-sourcing platform for the documentation of stem vocabulary in african sign languages. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 658–660, New York, NY, USA. Association for Computing Machinery.

Abdelhadi Soudi and Corinne Vinopol. 2019. Educational challenges for deaf and hard-of-hearing children in morocco. *Deaf education beyond the western world: Context, challenges, and prospects*, pages 307–322.

William C. Stokoe. 1960. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of Deaf Studies and Deaf Education*, 10(1):3–37.

Deborah Stryker, Diane Nielsen, and Barbara Luetke. 2015. Signing exact english: Providing a complete model of english for literacy growth.

Tony Veale, Alan Conway, and Bróna Collins. 1998. The challenges of cross-modal translation: English-to-sign-language translation in the zardoz system. *Machine Translation*, 13:81–106.

Tamara Wilson and Merv Hyde. 1997. The use of signed english pictures to facilitate reading comprehension by deaf students. *American Annals of the Deaf*, pages 333–341.

Liwei Zhao, Monica Costa, and NL Badler. 2000. Interpreting movement manner. In *Proceedings Computer Animation 2000*, pages 98–103. IEEE.

Harmonizing Annotation of Turkic Postverbial Constructions: A Comparative Study of UD Treebanks

Arofat Akhundjanova
Saarland University / Saarbrücken, Germany
arak00001@stud.uni-saarland.de

Abstract

As the number of treebanks within the same language family continues to grow, the importance of establishing consistent annotation practices has become increasingly evident. In this paper, we evaluate various approaches to annotating Turkic postverbial constructions across UD treebanks. Our comparative analysis reveals that none of the existing methods fully capture the unique semantic and syntactic characteristics of these complex constructions. This underscores the need to adopt a balanced approach that can achieve broad consensus and be implemented consistently across Turkic treebanks. By examining the phenomenon and the available annotation strategies, our study aims to improve the consistency of Turkic UD treebanks and enhance their utility for cross-linguistic research.

1 Introduction

As the Universal Dependencies (UD) project (Nivre et al., 2016, 2020) continues to grow, the need for consistent annotation practices across treebanks has become increasingly evident, especially for languages within the same language family. The Turkic language family, with its rich morpho-syntactic categories and agglutinative morphology, poses unique challenges for annotation. Despite the availability of several Turkic UD treebanks, inconsistencies in annotation schemes often hinder meaningful comparisons and cross-lingual studies, highlighting the necessity for a standardized approach.

Previous studies have emphasized inconsistencies in the annotation of Turkic languages, particularly in morphological features and dependency relations (Tyers et al., 2017). These include challenges in part-of-speech (POS) tagging, morphological features (Taguchi, 2022), and pronominalized locatives (Washington et al., 2024).

The development of the first UD treebank for Uzbek and the challenges faced during annotation prompted us to investigate a specific issue: the annotation of Turkic postverbial constructions. These constructions, which pair a converb with a postverb, convey nuanced meanings related to aspect or actionality. The dual role of postverbs — functioning both as grammatical markers and as independent verbal predicates — complicates their representation within the UD framework. Ensuring consistency while accurately reflecting the unique semantic and syntactic structure of postverbial constructions is difficult.

In this paper, we evaluate multiple approaches to annotating Turkic postverbial constructions across eleven UD treebanks of seven Turkic languages, as shown in Table 1. This issue is particularly critical given the variation not only across Turkic languages but also within the treebanks of a single language.

Our analyses and suggestions contribute to improving the consistency of Turkic UD treebanks and enhancing their value for cross-linguistic research.

The remainder of this paper is organized as follows. Section 2 provides background information on Turkic postverbial constructions. Section 3 presents a detailed analysis of four annotation approaches: `adverbial clause modifier`, `clausal complement`, `auxiliary` and `compound`. Section 4 offers recommendations for standardizing annotations, and Section 5 concludes our findings with implications for future work on Turkic UD treebanks.

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 10–17
January 20, 2025. ©2025 Association for Computational Linguistics

Treebanks	sent	tok	genre	No. of postverbial constructions
Azerbaijani-TueCL (Eslami and Çağrı Çöltekin, 2024)	109	663	grammar	∼ 4
Kazakh-KTB (Tyers and Washington, 2015)	1078	10536	news, fiction, wiki	∼ 24
Kyrgyz-KTMU (Benli, 2020)	2480	23654	news, fiction	∼ 60
Kyrgyz-TueCL (Chontaeva and Çağrı Çöltekin, 2024)	145	1001	grammar	∼ 30
Tatar-NMCTT (Taguchi et al., 2022)	148	2280	news, non-fiction	∼ 6
Turkish-BOUN (Türk et al., 2021)	9761	125212	news, non-fiction	∼ 100
Turkish-GB (Çağrı Çöltekin, 2015)	2880	17177	grammar	∼ 3
Turkish-Kenet (Kuzgun et al., 2022)	18687	178658	grammar	N/A
Turkish-Penn (Cesur et al., 2022)	16396	183555	news, non-fiction	N/A
Uyghur-UDT (Eli et al., 2016)	3456	40236	fiction	∼ 80
Uzbek-UT (Akhundjanova, 2024)	500	5850	news, fiction	∼ 70

Table 1: Eleven Turkic UD treebanks representing seven languages selected for our comparative study.

2 Turkic Postverbial Constructions

Turkic languages use verbal constructions made up of a converb followed by an auxiliary verb, also called a ‘postverb’ (Ağcagül, 2004) or ‘postverbial constructions with auxiliary verbs’ (Johanson, 2021, 36-37). In these constructions, the converb provides the main lexical meaning, while the postverb, having lost much of its original meaning, primarily carries grammatical information like person, mood and tense. It also refines the description of the action, as in Kyrgyz `kel-ip tur` (lit. ‘coming stand’), which means ‘to come regularly.’ The postverb adopts the converb’s argument structure, forming a single grammatical unit.

This structure bears similarity to Indo-European preverbal units, where a non-inflecting element precedes a verb stem, forming a unified lexical unit. Preverbs typically modify or refine the verb’s lexical meaning, adding spatial, directional, or aspectual nuances. For instance, in Sanskrit `pra gacchati` (lit. ‘forth goes’), the meaning is ‘he goes forth’ (Booij and Van Kemenade, 2003).

The following kinds of verbs can occur as the auxiliary element in postverb constructions of various Turkic languages: tur-/dur- ‘stand (up)’, yat-/yot-/jat- ‘lie (down)’, oltur-/otur-/o‘tir- ‘sit (down)’, kel-/kil-/gel- ‘come’, ket-/git- ‘go’, bar-/bor- ‘go’, al-/ol- ‘take’, ber-/bir-/ver- ‘give’, ïd-/yubor- ‘send’, etc (Ağcagül, 2004, 7).

Postverbs typically convey two types of functions:

1. Actional modification: Postverbs modify the actional meaning of the lexical verb by specifying qualitative or quantitative properties such as suddenness (1) and thoroughness (2) (Ağcagül, 2004, 7), as in the following examples:

(1) Uzbek

ayt-ib qo‘y-di-m
say-CONV put-PST-1SG

‘I blurted out’ (lit. ‘saying put’)

(2) Uyghur

Oq-up čïq!
read-CONV emerge.IMP

‘Read from beginning to end!’ (lit. ‘reading emerge’)

2. Phase specification: Postverbs indicate different phases of an action, including its initial or final stages, as well as its continuity (Ağcagül, 2004, 7), as illustrated in the examples below:

(3) Turkish

yaz-ıp dur-du
write-CONV stand-PST.3SG

‘s/he kept writing’ (lit. ‘writing standed’)

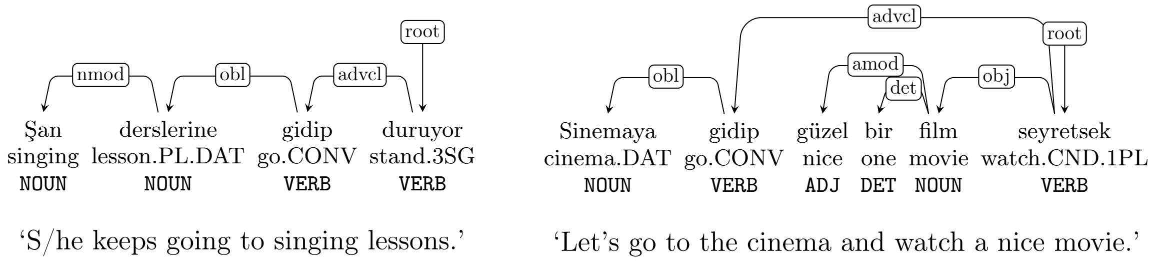
(4) Uzbek

Manzil-ga yet-ib qol-di-k
destination-DAT reach-CONV stay-PST-1PL

‘We are about to reach the destination.’ (lit. ‘destination.to reaching (we) stayed’)

3 Existing Annotation Approaches

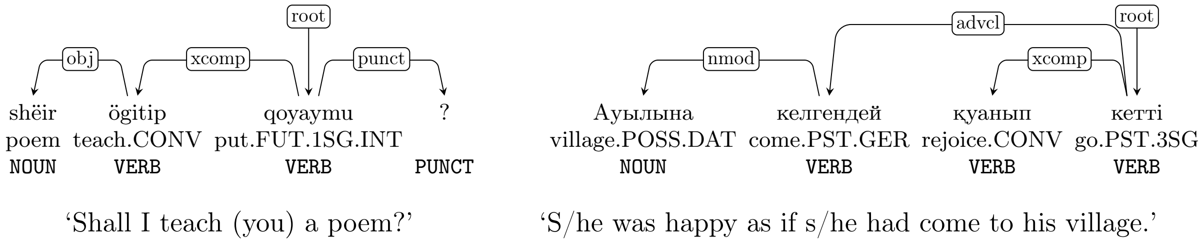
We examine four existing approaches to annotating Turkic postverb constructions, outlining the arguments for and against each. These approaches include treating them as `adverbial`



(a) Annotation for `gidip dur`.

(b) Annotation for `gidip`.

Figure 1: The converb `gidip` is used in two different structures, but tagged with the same label in the Turkish-GB treebank.



(a) Annotation for `ögitip qoy`.

(b) Annotation for `қуанып кет`.

Figure 2: Annotation of the converb as `xcomp` in Uyghur and Kazakh.

`clause modifier` (3.1), `clausal complement` (3.2), `auxiliary` (3.3), and `compound` (3.4). Additionally, we find instances of mixed approaches in certain treebanks (3.5).

3.1 Adverbial clause modifier: `advcl`

One approach to addressing this issue is to annotate the converb as `advcl` and the postverb as the `head`, as shown in Figure 1a. This method has been adopted in the Turkish treebanks listed in Table 1.

However, this annotation is not ideal. The `advcl` tag is generally reserved for clauses functioning as modifiers that express temporal, causal, conditional, or similar relations. In Turkic postverb constructions, the converb does not serve as a modifier to the postverb. Instead, it forms an integral part of the verbal phrase, contributing essential lexical meaning. Annotating the converb as `advcl` misrepresents its role, inaccurately suggesting that it has a subordinate function relative to the postverb. This approach fails to capture the grammaticalized and semantically unified nature of these constructions. For comparison, see Figure 1b, which shows a true adverbial clause modifier using the same converb `gidip`, contrasted with the postverbial construction in Figure 1a.

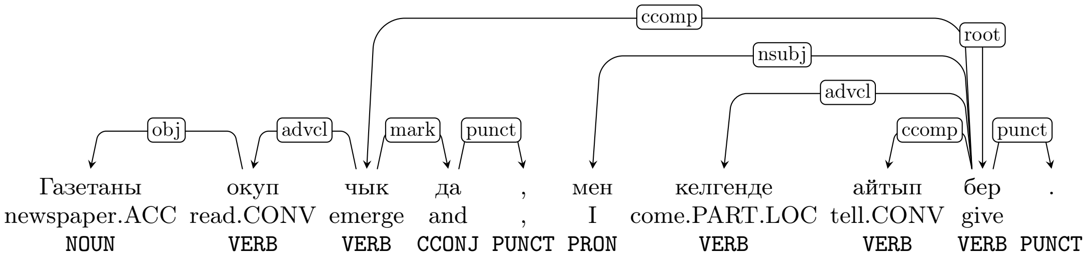
3.2 Clausal Complement: `xcomp` and `ccomp`

Another option is to tag the converb as `xcomp` (see Figure 2a for Uyghur and 2b for Kazakh) or `ccomp` (see Figure 3 for Kyrgyz) and the postverb as the head. This method is not plausible, because the two elements of postverbial constructions do not function as independent predicates, nor do they exhibit the syntactic independence typical of an `xcomp` or `ccomp` relation. In these relations, the complement clause is subordinate to the main predicate (head) and lacks its own subject, relying on an external argument for subject control. However, in postverbial constructions, the converb is not a subordinate clause but rather an integral part of a compound verb.

3.3 Auxiliary: `aux`

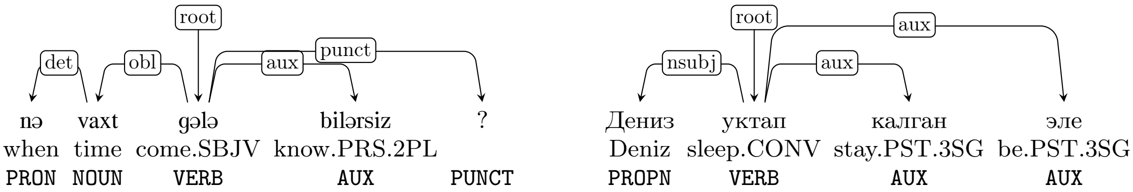
Tagging the converb as the head and the postverb as `aux` can be a reasonable approach in some contexts. See Figure 4a from Azerbaijani-TueCL, Figure 4b from Kyrgyz-TueCL and Figure 5 from Tatar-NMCTT. However, there are important considerations and potential limitations depending on the specific properties of the language.

On the one hand, the converb carries the primary lexical meaning, making it appropriate to treat it as the head. This reflects its domi-



‘Read the newspaper and tell me when I come.’

Figure 3: Annotation of `окуп чык` with `advcl` and `айтып бер` with `ccomp`.



‘When can you come?’

(a) Annotation for `gələ bil`.

‘Deniz had fallen asleep.’

(b) Annotation for уктап кал.

Figure 4: Annotation of the converb as a head and the postverb as `aux`.

nant role in encoding the core action or state of the clause. Postverbs are often grammaticalized to indicate auxiliary-like functions, which aligns with the typical `aux` tag. Treating the postverb as `aux` captures its secondary grammatical function and reduced lexical meaning. In both Azerbaijani and Kyrgyz treebanks, this approach is applied based on the classification of auxiliaries in their respective languages. In the Azerbaijani treebank, independent verbs like `bil` ‘know’ and `ol` ‘become’ are tagged with AUX POS, and Kyrgyz-TueCL treebank has a larger list of auxiliaries: жат, кал, ал, бол, кой, кет, тур, etc. Tatar treebank (Taguchi et al., 2022) also indicates that the finite verb in grammaticalized converb constructions is marked as AUX.

On the other hand, in other Turkic languages, postverbs often retain independent, non-auxiliary uses as lexical verbs and appear as heads of their own clauses with full argument structures. For example, compare the following two Uzbek sentences:

(5) Uzbek

yomg‘ir	qor-ni	eri-t-ib	yubor-di
rain	snow-ACC	melt-CAU-CONV	send-PST.3SG

‘The rain melted the snow away.’

(6)

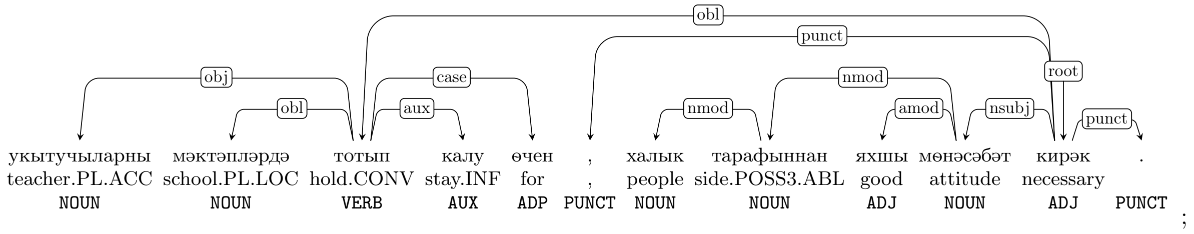
xat-ni	ber-ib	yubor-di
letter-ACC	give-CONV	send-PST.3SG

‘S/he gave/sent the letter away.’

In (5), the postverb `yubordi` ‘sent’ marks the immediate completion of the action expressed by the converb `eritib` ‘melting’. In (6), both `berib` ‘giving’ and `yubordi` ‘sent’ retain their independent meanings, and serve more like a serial verb construction (`compound:svc`). For this reason, in Uzbek, about 27 verbs that can be used as auxiliaries to form postverbial constructions are classified as `VERB`, not `AUX` and the `aux` relation is restricted to modal and copular verbs, and may not extend to aspectual or actionality markers. Hence, this approach would overload the `aux` with elements that do not fit its traditional definition.

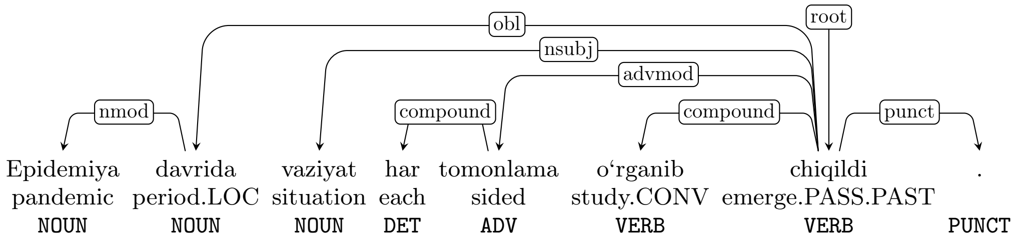
3.4 Compound

The final approach is to use a `compound` relation, as shown in the Uzbek-UT example in Figure 6. Postverb constructions are akin to compound verbs, where all elements contribute to forming a single lexical unit. However, we acknowledge that the `compound` label does not



'In order to retain teachers in schools, good public attitude is needed.'

Figure 5: Annotation of `тотып калу` as `aux`.



'During the epidemic, the situation was thoroughly studied.'

Figure 6: Annotation of `o'rganib chiq` as `compound`.

fully reflect the postverb's desemanticized and auxiliary-like role. Tagging the converb as `compound:lvc` (light verb construction, LVC) instead could be a partially plausible option. In such verbal constructions, the verbal or non-verbal predicate provides the main semantic content like converbs in our case, while the light verb contributes grammatical information, resembling postverbs. The `compound:lvc` relation highlights the grammaticalized nature and auxiliary function of the postverb while still acknowledging the converb as the core semantic contributor. It aligns with the principle that LVCs combine a semantically strong element with a semantically weak verb.

The limitation of this approach is that Turkic postverb constructions are highly grammaticalized, often to the point where the postverb functions more like an auxiliary than a light verb. As a result, using the `compound:lvc` might not fully capture this advanced stage of grammaticalization.

3.5 Mixed Approaches

The inconsistency in annotation methods within the same language or treebank may stem from several factors.

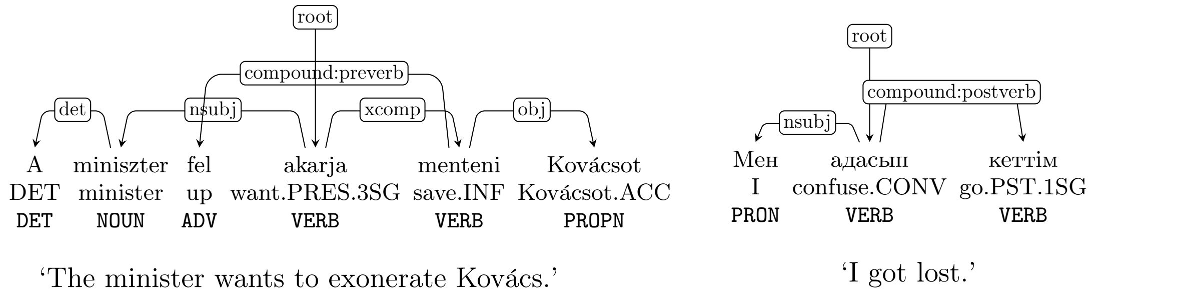
Firstly, distinguishing postverbial constructions from superficially similar multiverb constructions can be challenging. This often involves determining whether the second verb functions as a lexical verb or an auxiliary. For instance, as illustrated in (6), verbs like `yubor` may carry the lexical meaning 'to send' or modify an actional content, as in 'to do immediately and easily.' So, the phrase `ber-ib yubor` (give-CONV send) can be interpreted either as the lexical action 'to send,' i.e., 'to send through someone,' or as an actional modification of `ber` ('to give'), meaning 'to give immediately.'

Secondly, combinations of postverbial constructions can further complicate analysis. For example, in Uzbek, the phrase `yoz-ib ber-a qol` (write-CONV give-CONV remain) combines `yoz-ib ber` ('to write for someone') with `qol` ('to remain') to mean 'to start writing for someone' (Kononov, 1960, 268).

Such ambiguities significantly complicate both analysis and annotation. For instance, in the Kyrgyz-KTMU treebank, two postverbial constructions within the same sentence are analyzed differently. As shown in Figure 3, `оку-п чык` (read-CONV emerge) 'to read thoroughly' is annotated with the `advcl` relation, whereas `айт-ып бер` (tell-CONV give) 'to tell somebody' is annotated with the `ccomp` relation. Similar inconsistencies are also observed in several Turkish treebanks.

Approach	Treebank	Head Type	Cross-linguistic Applicability	Compliance with UD Guidelines	Frequency in Treebanks
advcl	Turkish-BOUN Turkish-Penn Turkish-Kenet Turkish-GB Kyrgyz-KTMU	postverb	no	no	high
xcomp/ccomp	Uyghur-UDT Kazakh-KTB	postverb	no	no	medium
aux	Azerbaijani-TueCL Kyrgyz-TueCL Tatar-NMCTT	converb	yes	yes	low
compound	Uzbek-UT	postverb	yes	yes	low

Table 2: Summary of annotation approaches for Turkic postverb constructions, detailing head type, cross-lingual applicability, compliance with UD guidelines, and the frequency of each approach across treebanks.



(a) Annotation for a Hungarian preverbal construction.

(b) Annotation for a Kazakh postverbial construction.

Figure 7: Possible annotation of a postverbial construction using `compound:postverb`, analogous to `compound:preverb`.

4 Discussion

The summary of the approaches described in Section 3 with their advantages and disadvantages is given in Table 2.

Tagging the converb as an adverbial clause or clausal complement while assigning the postverb as the head misrepresents the tight syntactic and semantic integration of Turkic postverb constructions. Although these two methods highlight that the converb conveys the primary lexical meaning, and are relatively common among Turkic treebanks, they do not fully adhere to UD guidelines or cross-linguistic annotation practices.

Tagging the converb as the head and the postverb as `aux` can be a reasonable approach in some contexts. In many languages, auxiliaries are desemanticized elements that support the main verb. This pattern can apply to Turkic postverbs when they primarily serve grammatical functions. However, in some Turkic languages, they might retain sufficient lexical meaning or syntactic independence to argue against classifying them as auxiliaries. For instance, if postverbs retain a significant degree of lexical meaning, a different relation such as `compound:lvc` or `compound:svc` might be more accurate.

Each of these methods has its strengths and limitations. A potential alternative could be to introduce a new language-specific subtype relation, such as `compound:postverb`, mirroring the logic behind the `compound:preverb` relation used in the Hungarian treebank (Vincze et al., 2010). This approach would avoid the misapplication of generic relations like `compound`. Figure 7 illustrates the proposed `compound:postverb` relation alongside the Hungarian example annotated with `compound:preverb`.

5 Concluding Remarks

We agree that the best approach to annotating Turkic postverbial constructions depends on the specific properties of the language and the constraints of the annotation framework. Based on our analysis, the `compound` approach seems to be the most suitable, but we propose a dedicated subtype, `compound:postverb`, to balance semantic accuracy, syntactic clarity, and cross-linguistic comparability within the UD framework. We emphasize the importance of collaborative discussions among UD contributors, including cross-lingual and cross-treebank exchanges, to ensure robust annotation guidelines. In the future, we plan to organize a shared task within a UD Working Group to identify the optimal solution and validate the proposed annotation approach. Consistent tagging across languages and treebanks will strengthen the universality of UD, support typological linguistic studies, and foster cross-lingual applications in natural language processing (NLP).

References

Arofat Akhundjanova. 2024. UD Uzbek UT. https://github.com/UniversalDependencies/UD_Uzbek-UT.

Sevgi Ağcagül. 2004. Grammaticalization of turkic postverbial construction. Orientalia Suecana, 53:5–14.

Ibrahim Benli. 2020. UD Kyrgyz KTMU. https://github.com/UniversalDependencies/UD_Kyrgyz-KTMU.

Geert Booij and Ans Van Kemenade. 2003. Preverbs: an introduction, pages 1–11. Springer Netherlands, Dordrecht.

Neslihan Cesur, Aslı Kuzgun, Olcay Taner Yıldız, Büşra Marşan, Neslihan Kara, Bilge Nas Arıcan, Merve Özçelik, and Deniz Baran Aslan. 2022. UD Turkish Penn. https://github.com/UniversalDependencies/UD_Turkish-Penn.

Bermet Chontaeva and Çağrı Çöltekin. 2024. UD Kyrgyz TueCL. https://github.com/UniversalDependencies/UD_Kyrgyz-TueCL.

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.

Soudabeh Eslami and Çağrı Çöltekin. 2024. UD Azerbaijani TueCL. https://github.com/UniversalDependencies/UD_Azerbaijani-TueCL.

Lars Johanson. 2021. The Structure of Turkic. In Lars Johanson and Éva Á. Csató, editors, The Turkic languages, pages 26–59. Routledge.

A.N. Kononov. 1960. Grammatika sovremennogo uzbekskogo literaturnogo jazyka [Grammar of the Modern Uzbek Literary Language]. Moskva: Izdatel'stvo Akademii Nauk SSSR.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Arife Betül Yenice, Bilge Nas Arıcan, and Ezgi Sanıyar. 2022. UD Turkish Kenet. https://github.com/UniversalDependencies/UD_Turkish-Kenet.

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).

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.

Chihiro Taguchi. 2022. Consistent grammatical annotation of Turkic languages for more universal Universal Dependencies. In 29th International Conference on Head-Driven Phrase Structure Grammar.

Chihiro Taguchi, Sei Iwata, and Taro Watanabe. 2022. Universal Dependencies treebank for Tatar: Incorporating intra-word code-switching information. In Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference, pages 95–104, Marseille, France. European Language Resources Association.

Francis Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An assessment of universal dependency annotation guidelines for turkic languages. In Proceedings of the 5th International Conference on Turkic Languages Processing (TurkLang 2017), pages 276–297.

Francis M. Tyers and Jonathan N. Washington. 2015. Towards a free/open-source universal-dependency treebank for kazakh. In 3rd International Conference on Turkic Languages Processing, (TurkLang 2015), pages 276–289.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2021. Resources for turkish dependency parsing: Introducing the boun treebank and the boat annotation tool. Preprint, arXiv:2002.10416.

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).

Jonathan Washington, Çağrı Çöltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Gulnura Jumalieva, Aida Kasieva, Aslı Kuzgun, Büşra Marşan, and Chihiro Taguchi. 2024. Strategies for the annotation of pronominalised locatives in Turkic Universal Dependency treebanks. In Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, pages 207–219, Torino, Italia. ELRA and ICCL.

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.

Towards Truly Open, Language-Specific, Safe, Factual, and Specialized Large Language Models

Preslav Nakov
Mohamed bin Zayed University of Artificial Intelligence
Masdar City, Building 1B, 3rd Floor
Abu Dhabi, UAE

Abstract

First, we will argue for the need for fully transparent open-source large language models (LLMs), and we will describe the efforts of MBZUAI's Institute on Foundation Models (IFM) towards that based on the LLM360 initiative. Second, we will argue for the need for language-specific LLMs, and we will share our experience from building Jais, the world's leading open Arabic-centric foundation and instruction-tuned large language model, Nanda, our recently released open Hindi LLM, and some other models. Third, we will argue for the need for safe LLMs, and we will present Do-Not-Answer, a dataset for evaluating the guardrails of LLMs, which is at the core of the safety mechanisms of our LLMs. Forth, we will argue for the need for factual LLMs, we will discuss the factuality challenges that LLMs pose. We will then present some recent relevant tools for addressing these challenges developed at MBZUAI: (i) OpenFactCheck, a framework for fact-checking LLM output, for building customized fact-checking systems, and for benchmarking LLMs for factuality, (ii) LM-Polygraph, a tool for predicting an LLM's uncertainty in its output using cheap and fast uncertainty quantification techniques, and (iii) LLM-DetectAIve, a tool for machine-generated text detection. Finally, we will argue for the need for specialized models, and we will present the zoo of LLMs currently being developed at MBZUAI's IFM.

Bio:

Bio

Preslav Nakov is Professor and Department Chair for NLP at the Mohamed bin Zayed University of Artificial Intelligence. He is part of the core team at MBZUAI's Institute for Foundation Models that developed Jais, the world's best open-source Arabic-centric LLM, Nanda, the world's best Hindi model, and LLM360, the first truly open LLM. Previously, he was Principal Scientist at the Qatar Computing Research Institute, HBKU, where he led the Tanbih mega-project, developed in collaboration with MIT, which aims to limit the impact of "fake news", propaganda and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking. He received his PhD degree in Computer Science from the University of California at Berkeley, supported by a Fulbright grant. He is Chair-Elect of the European Chapter of the Association for Computational Linguistics (EACL), Secretary of ACL SIGSLAV, and Secretary of the Truth and Trust Online board of trustees. Formerly, he was PC chair of ACL 2022, and President of ACL SIGLEX. He is also member of the editorial board of several journals including Computational Linguistics, TACL, ACM TOIS, IEEE TASL, IEEE TAC, CS&L, NLE, AI Communications, and Frontiers in AI. He authored a Morgan & Claypool book on Semantic Relations between Nominals, two books on computer algorithms, and 250+ research papers. He received a Best Paper Award at ACM WebSci'2022, a Best Long Paper Award at CIKM'2020, a Best Resource Paper Award at EACL'2024, a Best Demo Paper Award (Honorable Mention) at ACL'2020, a Best Task Paper Award (Honorable Mention) at SemEval'2020, a Best Poster Award at SocInfo'2019, and the Young Researcher Award at RANLP'2011. He was also the first to receive the Bulgarian President's John Atanasoff award, named after the inventor of the first automatic electronic digital computer. His research was featured by over 100 news outlets, including Reuters, Forbes, Financial Times, CNN, Boston Globe, Aljazeera, DefenseOne, Business Insider, MIT Technology Review, Science Daily, Popular Science, Fast Company, The Register, WIRED, and Engadget, among others.

Proceedings of the 18th Building and Using Comparable Corpora Workshop, page 18
January 20, 2025. ©2025 Association for Computational Linguistics

Make Satire Boring Again: Reducing Stylistic Bias of Satirical Corpus by Utilizing Generative LLMs

Asli Umay Ozturk[1], Recep Firat Cekinel[1], Pinar Karagoz[1]

[1] Department of Computer Engineering
Middle East Technical University (METU)
{auozturk,rfcekinel,karagoz}@ceng.metu.edu.tr
Correspondence: auozturk@ceng.metu.edu.tr

Abstract

Satire detection is essential for accurately extracting opinions from textual data and combating misinformation online. However, the lack of diverse corpora for satire leads to the problem of stylistic bias which impacts the models' detection performances. This study proposes a debiasing approach for satire detection, focusing on reducing biases in training data by utilizing generative large language models. The approach is evaluated in both cross-domain (irony detection) and cross-lingual (English) settings. Results show that the debiasing method enhances the robustness and generalizability of the models for satire and irony detection tasks in Turkish and English. However, its impact on causal language models, such as Llama-3.1, is limited. Additionally, this work curates and presents the Turkish Satirical News Dataset with detailed human annotations, with case studies on classification, debiasing, and explainability.

1 Introduction

There are no universally agreed definitions for satire, sarcasm, and irony in NLP literature. However, Cambridge Dictionary[1] defines **satire** as a way of *criticizing* people or ideas in a *humorous* way, especially in order to make a political point. Whereas **sarcasm** is defined as the use of remarks that clearly *mean the opposite* of what they say, made in order to hurt someone's feelings or to *criticize* something in a *humorous* way. Moreover, **irony** is related to the use of words that are the *opposite of what you mean*, as a way of being *funny*. The overlaps in these definitions cause different studies to use these terms interchangeably, and borrow ideas from other studies (Barbieri et al., 2014; Van Hee et al., 2016a, 2018; Carvalho et al., 2020).

At first glance, it might not be very apparent why detecting satirical content is important. However, with the increased usage of social media, the primary source of news and information for many people has become the shared news articles in their social media feeds. Even though this makes the information more accessible, it can also cause misinformation to spread at fast rates (Allcott and Gentzkow, 2017; Aïmeur et al., 2023). It is not uncommon for regular social media users to take fake or satirical content as the truth (Wu et al., 2019), which is specifically problematic when it comes to news content. Hence, satire detection can offer a solution to this misinformation problem since automated detection of satirical content can be used to create automated warnings that inform social media users about the reliability of a piece of information.

For the last couple of years, with the rise of LLMs, an improvement in the performance of NLP tasks has been seen in the literature (Li et al., 2022). Even though recent studies have demonstrated high performances for satire, sarcasm, and irony classification tasks using multilingual LLMs, it remains unclear whether these concepts are represented similarly across different languages and domains (Ortega-Bueno et al., 2023; Maladry et al., 2023).

Moreover, focusing on a data-centric approach when training LLMs also raises problems since it is not easy to find a diverse set of resources for low-resourced languages (Doğru et al., 2018; Hangya et al., 2022; Acikgoz et al., 2024) or specific tasks. For example, annotating data that can be labeled as satirical, sarcastic, or ironic requires extensive human labor. Instead, automatic data collection processes may be employed, and the data would be collected from a limited number of sources that are already known to belong to the target label. As a result, this creates stylistically unbalanced corpora to be used to train and fine-tune LLMs. This may result in a bias or misalignment in the model (Xu et al., 2024). In other words, potential stylistic bias

[1] https://dictionary.cambridge.org/dictionary/english

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 19–35
January 20, 2025. ©2025 Association for Computational Linguistics

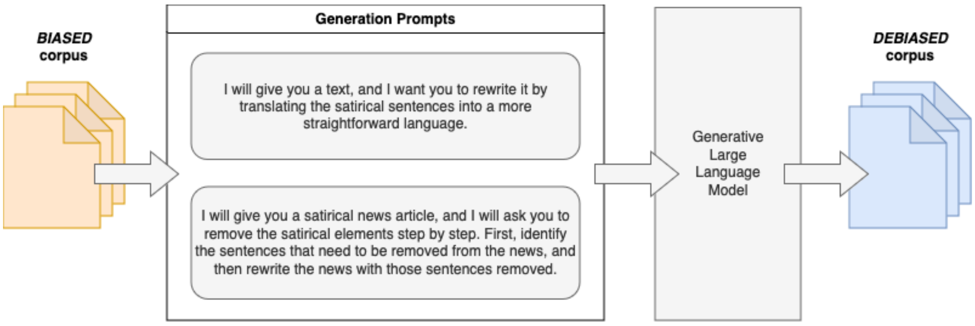

BIASED corpus
Generation Prompts
I will give you a text, and I want you to rewrite it by translating the satirical sentences into a more straightforward language.
I will give you a satirical news article, and I will ask you to remove the satirical elements step by step. First, identify the sentences that need to be removed from the news, and then rewrite the news with those sentences removed.
Generative Large Language Model
DEBIASED corpus

Figure 1: The proposed debiasing pipeline

in the curated dataset impacts the robustness of the models that use these datasets during training and fine-tuning.

This work aims to reduce the effect of stylistic bias stemming from a single-sourced satirical corpus, proposing a debiasing method that utilizes generative LLMs to reduce the stylistic bias of the instances in the biased corpus.

The proposed method works by generating satirical texts that are stylistically more natural, making the generated corpus more parallel to the non-satirical corpus. This method is demonstrated on a curated dataset for satirical news detection in Turkish. Our contributions can be summarized as follows:

- Curating the *Turkish Satirical News Dataset* with human annotations, and analyzing its stylistic bias and usability[2],
- Proposing a *debiasing pipeline (Figure 1)* for combating stylistic bias and improving model generalizability,
- Analyzing the cross-lingual and cross-domain performance of Turkish satire detection model for irony and English.

2 Related Work

Starting in the early 2010s, the related literature began to focus on the problems of binary detection for satire, sarcasm, and irony. Earlier works generally utilize traditional supervised learning methods such as Support Vector Machine (SVM) or Naive Bayes (NB) based classifiers and propose different feature extraction methods for different languages and tasks (Buschmeier et al., 2014; Barbieri et al., 2014; Van Hee et al., 2016b; Pamungkas and Patti, 2018; Baloglu et al., 2019; Ozturk et al., 2021; Onan and Toçoğlu, 2020). Another approach explored in the earlier studies is utilizing neural network based architectures such as LSTM (Long-Short Term Mermory) networks (Wu et al., 2018; Zhang et al., 2019a). Later works started to utilize transformer architectures such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019a) and reported improved results (Buyukbas et al., 2021).

In the last couple of years, newer studies have focused more and more on multimodal approaches and LLM-based models. In their study, Tomás et al. (2023) explore the irony detection performance of transformer-based models with both textual and visual inputs. On the other hand, Lin et al. (2024) combines transformer-based models with prompt engineering to improve the irony detection performance, specifically focusing on different features of the text.

A recurring problem in the literature for the aforementioned tasks is the lack of labeled data and openly available datasets. There are curated datasets for irony detection (Van Hee et al., 2016a) and sarcastic news detection (Barbieri et al., 2014) but they are mostly in English. Even though some other datasets curated for other languages do exist (Ortega-Bueno et al., 2019; Xiang et al., 2020; Ozturk et al., 2021; Ghanem et al., 2020; Joshi et al., 2017), they are much smaller than the available English corpus for irony, sarcasm, and satire.

Literature on irony, satire, and sarcasm detection also includes studies utilizing explainable artificial intelligence (AI) and interpretable machine learning (ML) methods (Buyukbas et al., 2021) such as LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) and SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) or proposing task-based structures to understand model decisions (Shu et al., 2019). These studies aim to analyze the model decisions to improve the performance, fairness, and generalizability of the

[2] https://github.com/auotomaton/satiretr

model, as well as reduce the bias of the model.

With the advancements in generative LLMs, there have been studies utilizing them to augment training data or synthetically generate data from scratch, to overcome the shortness of training data available for a diverse set of tasks (Hangya et al., 2022; Long et al., 2024).

A couple of works have tried to utilize this data generation approach to overcome biases in datasets. Qian et al. (2021) focus on dataset bias and propose a framework for debiasing using counterfactual inference. They show that their approach improves the effectiveness, generalisability, and fairness of the classifier. In another study, Schlicht et al. (2024) utilize conversational LLMs to reduce textual bias in news articles. Their findings show that even though they are compelling in some cases, they tend to leave out vital and contextual information during the debiasing process.

Differing from the existing works, this work explores the effects of synthetically generated data on debiasing binary classifiers trained on low-resource languages, focusing on satire detection in Turkish. We propose a debiasing pipeline that aims to improve the cross-lingual and cross-domain performance of a model trained on a stylistically biased dataset.

3 Dataset

One contribution of this study is to curate an open dataset for satire and satirical news detection in Turkish. Utilizing the satirical news publication Zaytung[3] and Turkish news agency Anadolu Agency (AA)[4], *Turkish Satirical News Dataset* is curated.

3.1 Curation of Turkish Satirical News Dataset

As a source of satirical news articles, the Turkish satirical news publication Zaytung is used. By crawling the Zaytung website archive, 2825 satirical articles are collected with *timestamp*, *title*, *body* and *header image* information. To improve the representative nature of the dataset, articles dated before 2014 are discarded.

As a source of non-satirical news articles, AA archives are crawled between the dates 2022-2023, and 4781 articles are collected with *id*, *title*, *subtitle*, *author*, *date*, *category*, *header image*, *city*, *body* and *tag* information.

The final dataset includes 2202 SATIRICAL and 4781 NON-SATIRICAL articles. Additionally, 40 of the SATIRICAL instances have word-by-word human annotations aiming to capture the effect of each word on the satirical meaning of the article. The annotation process and a case study comparing the annotations with model explanations are presented in Appendix A and B.

The curated dataset is available publicly on GitHub[5] for other researchers' use. The GitHub repository also includes the debiased articles generated from the Zaytung corpus, described in Section 4. All data is scraped from publically available sources and Zaytung has specifically been reached for consent to use their articles.

3.2 Bias Analysis

Since both the SATIRICAL and the NON-SATIRICAL corpora are taken from a single publication each, it might be expected to see a stylistic and statistical bias that results in an easily separable dataset. This may result in classifiers that are trained using such datasets to become biased for the style elements of the corpus instead of identifying satire/non-satire indications of the text. One simple way to observe this is by conducting statistical analysis on both the SATIRICAL and NON-SATIRICAL corpora.

3.2.1 Average Word and Sentence Count

A primary statistical analysis is performed and reported in Table 1 to better understand the data instances. It can be seen that SATIRICAL corpus has an average of 329 words per instance and 44 sentences per instance. On the other hand, NON-SATIRICAL corpus has an average of 313 words per instance and 43 sentences per instance. Even though the numbers are close, on average, we see that the SATIRICAL corpus has more words per sentence.

Statistic	SATIRICAL	NON-SATIRICAL
avg #of words	329	313
avg #of sentences	44	43

Table 1: Statistics of the *Turkish Satirical News Dataset* by corpus

3.2.2 Top 10 Words

To have a general idea about the content of the news belonging to both the SATIRICAL and NON-SATIRICAL corpora, top-10 terms are extracted per label by TF-IDF (Term Frequency - In-

[3] https://zaytung.com/

[4] https://www.aa.com.tr/

[5] https://github.com/auotomaton/satiretr

Label	Top 10 Words
SATIRICAL	*almak (take), bir (one/a), demek (say), etmek (make), gelmek (come), iş (work/job), olarak (being), vermek (give), türkiye (Turkiye), yapmak (do)*
NON-SATIRICAL	*ülke (country), yıl (year), açıklama (explanation), ifade (expression), fotoğraf (photograph), spor (sport), bölge (region), başkan (president), konu (issue), çalışmak (work)*

Table 2: Statistics of the *Turkish Satirical News Dataset* by Top 10 Words

verse Document Frequency) scoring. These terms are shown in Table 2. It is visible that the top 10 words for the SATIRICAL and NON-SATIRICAL corpora do not have any words in common. This also follows the idea that the tones of the two corpora are different.

4 The Proposed Debiasing Method

Bias introduced from the style of the source of the corpus is a serious concern that is hard to eliminate without extensive human annotation. This is specifically prevalent in fairly low-resource languages (Shen et al., 2024) and text qualities such as satirical, sarcastic, or ironic meaning (Maladry et al., 2023; Ortega-Bueno et al., 2023). Using a single source as a corpus makes the models more prone to bias. Hence, this study proposes the *debiasing pipeline* utilizing LLMs for generating less biased counterparts of texts.

4.1 Pipeline Design

The *debiasing pipeline* proposed in this work utilizes prompt engineering and synthetic data generation to remove the effect of the bias coming from the heavily stylistic language of the SATIRICAL corpus. The proposed pipeline is created to improve the usability of the curated dataset, however, we believe the pipeline can be generalized for any biased dataset by adapting the generation prompts according to the task and the bias.

The *debiasing pipeline*, summarised in Figure 1, generates a new set of data to be used to replace the training instances from the biased corpus. In the scope of this work, the stylistically biased data in the train set is the SATIRICAL class coming from the scraped Zaytung corpus.

4.2 Generating "Debiased" Articles Using Prompt Engineering

This subsection goes over the prompt engineering and debiased instance generation using sample articles translated so that it will be easier for readers to follow. The original Turkish articles and Turkish prompts are available in Appendix C. In this study, the data generation is done via the GPT web interface, ChatGPT[6]. However, the generation pipeline can easily be automatized by using the OpenAI API to fully automate the debiasing process. The following two prompts are used to generate stylistically less biased articles:

Prompt 1: *"I will give you a satirical news article, and I will ask you to remove the satirical elements step by step. First, identify the sentences that need to be removed from the news, and then rewrite the news with those sentences removed. Article text:"*

Prompt 2: *"I will give you a text, and I want you to rewrite it by translating the satirical sentences into a more straightforward language. Article text:"*

The first prompt, Prompt 1, asks the generative LLM to identify the satirical elements in the text and rewrite the article by excluding them. On the other hand, Prompt 2 asks the generative LLM to directly rewrite the article by modifying the sentences with satirical meaning to have a more straightforward language. Both prompts put the generative LLM to test in terms of its language understanding.

First, consider Sample Article 1 and the article generated from it using Prompt 1, shown in Figure 2. Even though the prompt clearly asks for the removal of satirical elements, the satire created by the fake social media expectation narrative is still present in the generated article. However, we see that the style of the writing is less exaggerated. This makes Prompt 1 a good candidate for generation in some cases.

On the other hand, consider Sample Article 2 and the articles generated from it using both Prompt 1 and Prompt 2, shown in Figure 3. It can be seen that Prompt 1 removes most of the sentences contributing to the satire in the text which significantly reduces the satirical meaning of the overall article. Without background knowledge about the financial situation revolving around Turkish Lira (TL), the article becomes vaguely satirical or even non-satirical to the uninformed reader. However, the

[6] https://chatgpt.com/

NASA Announces Termination of Mars Program Due to Photos Not Reaching Expected Like Counts. . .
The American Space Agency NASA is facing troubled days due to the photos sent by the Perseverance rover following its landing on Mars.
NASA's one-ton Rover-type spacecraft, Perseverance, landed successfully on Mars' Jezero Crater at 15:55 Eastern Time on Thursday after a journey of approximately 7 months. However, 24 hours after the landing, when NASA officials shared the first photos from the red planet on their Instagram account, they expressed surprise and disappointment over the comments they received. The photos, which were expected to garner at least 10 million likes, only managed to accumulate 2 million, leading to significant disillusionment within the community.
Comments such as "Did you send an iron to Mars to take these photos?", "You'd think they'd put a decent camera on it", and "There was no need to go all the way to Mars for this, we could have sorted it out in Yozgat city center" were particularly hurtful, according to James Watzin, Director of NASA's Mars Program.
Watzin admitted, "What's truly upsetting is that our followers are absolutely right. Honestly, we weren't satisfied either. If it wasn't our own rover, we wouldn't have liked those photos. It's just not appealing. . . "
Watzin pointed out that the photos failed to receive as many likes as a makeup-free selfie by Cardi B, noting, "If we don't receive something like a group of Martian mating photos in the next few days, it seems impossible to reach the number of likes needed to cover the program's costs. The effort we put in seems to have missed its mark. Why did this happen? Did we choose the wrong cameras? Is Mars itself not photogenic? These are questions that need answering." He announced that the Mars program will be put on hold until the source of the issue is understood.

(a) The original article, translated to English

The American Space and Aeronautics Agency NASA is going through difficult days due to the photos sent after the Perseverance rover successfully landed on Mars. Contrary to expectations, NASA officials encountered low engagement on social media following the successful landing of the Mars program. Program director James Watzin admitted that the lack of expected interest has led to disappointment and that the program has been deemed a failure. The Mars program has been paused to understand the source of the problem.

(b) Generated article, translated to English

Figure 2: Debiasing of Sample Article (1) with Prompt 1

The Crisis in LUNA is Growing: It Has Been Revealed to Have Lost Value Even Against the Turkish Lira...
Following a 99.8% loss in value, where Terra Luna Coin dropped from $64 to $0.20 within two days, bad news continues to emerge. After this dramatic drop that shook the cryptocurrency market, Do Kwon, CEO of Terraform Labs, released a statement today. He noted that their investigation revealed Luna had lost value not only against the US dollar but also against the Turkish Lira. Kwon said, "Everything we thought couldn't happen in the last two days has happened. I am very sorry."
Kwon explained that the realization came during a meeting he held to boost the team's morale, saying, "I gathered the team to give a motivational speech along the lines of 'Look, it's not so bad. At least we are still gaining value in TL terms.' I hadn't felt the need to check if we were worse off compared to TL. That was the real mistake." He shared that the painful truth became apparent when they opened the graph on the screen.
"At that moment, half the team left the building. We still can't reach them. The remaining team members are just looking around the office to see if there's anything left to sell," Kwon described the current situation at Terra Labs. The experienced CEO also mentioned that they expect the depreciation of the TL to eventually surpass Luna, whether due to some absurd move by Turkey or the Central Bank of the Republic of Turkey depleting its remaining foreign currency reserves. He added that he has not neglected to give optimistic messages for the future.

(a) The original satirical article, translated to English

Terra Luna Coin has experienced a significant loss in value in recent days. Terraform Labs CEO Do Kwon stated that their investigation revealed that Luna has lost value not only against the US dollar but also against the Turkish Lira (TL). This situation has caused a morale decline among company employees. Kwon expressed that in the short term, they expect the depreciation of the TL to surpass that of Luna.

(b) Generated article with Prompt 1, translated to English

In the past two days, Terra Luna Coin has experienced a 99.8% loss in value, dropping from $64 to $0.2, and bad news continues to emerge. Following this major drop, Terraform Labs CEO Do Kwon stated in a recent announcement that they have found evidence that Luna has lost value not only against the US dollar but also against the Turkish Lira. Kwon said, "Everything we didn't expect to happen in these last two days happened. I'm very sorry."
Kwon mentioned that they realized the situation during a meeting he organized to boost the team's morale. "I wanted to tell the team, 'At least we're gaining value against the TL,' but I hadn't checked the situation. That was the real mistake," he said. He also mentioned that they faced the harsh reality when they opened the chart.
"During the meeting, half of the team left the building, and we still can't reach them. The rest are just looking for things to sell in the office," Kwon said. He also noted that in the short term, the loss of value of the TL might surpass that of Luna, and he did not forget to provide an optimistic message for the future.

(c) Generated article with Prompt 2, translated to English

Figure 3: Debiasing of Sample Article (2) with Prompt 1 and Prompt 2

generated article with Prompt 2 can still be classified as satirical easily, while stripped from some of the more dramatic and stylistically strong phrases (e.g. *"...absurd move by Turkey or the Central Bank of the Republic of Turkey..."* etc.). This makes Prompt 2 also a good candidate for generation.

Another successful example that shows an original-generated article pair generated using Prompt 2 can be seen in Appendix C. With these explorations done on the prompts and their performances, both prompts are used during the generation process of 200 debiased articles.

4.3 Comparing Generated and Original Articles

The proposed debiasing method is expected to keep the satirical value of the articles while making them stylistically less biased. To check if this is the case, two analyses are conducted.

Firstly, the generated articles are checked manually (by one annotator) to see if they still can be classified as SATIRICAL. Additionally, they are analyzed to see if the original context of the article is still understandable or if some of the context (such as the events, relationships between people, and such) has been lost during the generation process. It is seen that:

- Out of the 200 articles, 29 of them can be labeled as NON-SATIRICAL by an unsuspecting reader with not enough knowledge of the Turkish political landscape. [7]

[7] Authors are aware that this definition of such a reader is highly subjective. This statistic is obtained to have a surface-level understanding of the quality of the generated content.

- Out of the aforementioned 29 articles, all 29 of them differ from their original counterparts with a loss of important contextual information, such as the total erasure of events or people occurring during the generation process.

- Out of the aforementioned 29 articles, 28 of them are generated with Prompt 1. Since Prompt 1 explicitly asks the model to remove satirical sentences, it is understandable to see a major loss in contextual information. This also points to Prompt 2 as being a better option for article generation.

As a second way of verifying whether the content of the articles in the corpus is maintained, the BERTScore (Zhang et al., 2019b) evaluation metric, which calculates the pairwise semantic similarity of tokens in the given pair of sentences using BERT's contextual embeddings, is employed. We used the BERTurk model for extracting contextual embeddings and the cosine similarity of the original sarcastic news and the debiased counterparts is 0.6852 (in terms of F1-binary). This similarity score implies that the debiased text mostly retains the original content.

5 Experiments

The experiments were conducted on 4 Nvidia A6000 GPUs. We employed three training strategies to evaluate the proposed debiasing pipeline's effectiveness.

5.1 Models and Parameters

We employed multilingual BERT (Devlin et al., 2019b), BERTurk (Schweter, 2020), XLM-Roberta-large (Conneau et al., 2020) and Llama-3.1-8B-Instruct (Dubey et al., 2024) models for conducting the experiments. These models were mainly selected because they were also pre-trained in Turkish. Llama-3.1-8B is a generative causal language model and we used sampling-based decoding to make predictions. As a result, for some test cases, the model's inferences do not comply with the expected labels and we reported the rate of nonresponses in Table 3. The samples where the model failed to return any labels were excluded during the evaluation of the models on the test sets.

During the training, we randomly selected 10% of the training data for validation. A grid search was conducted to explore the following hyperparameters: *learning rate* 2e-5, 5e-5, 1e-4 and *batch size* 8, 16, 32. The best results, as determined through this process, are reported in Section 5.3. All models were trained on the training datasets for two epochs.

Dataset	COMBINED	BIASED	DEBIASED
Zaytung	0.009	0.002	0.000
Onion	0.013	0.207	0.030
IronyTR	0.057	0.012	0.002

Table 3: Rate of nonresponses of Llama-3.1-8B-Instruct model

For the Llama-3.1 model, we set the sequence length to 2048, while for other models, the maximum sequence length was set to 512. Additionally, we employed QLoRA (Dettmers et al., 2024) for training the Llama model, with the LoRA rank and LoRA alpha set to 32 and the LoRA dropout to 0.05. We used the Adam optimizer with a cosine scheduler and trained the models with fp16 precision.

5.2 Training Setups and Test Datasets

The models were trained on three different setups as follows (illustrated in Figure 4):

BIASED: 200 instances from the SATIRICAL corpus and 200 instances from the NON-SATIRICAL corpus in *Turkish Satirical News Dataset* (detailed in Section 3) are selected.

DEBIASED: Same sets of instances as the *BIASED* setup are selected first, then the selected SATIRICAL instances are passed through the proposed *debiasing pipeline*. Final *DEBIASED* setup includes 200 instances from the SATIRICAL corpora and 200 instances from the NON-SATIRICAL corpora, where the SATIRICAL instances are debiased stylistically.

HYBRID: To have an intermediate setup between *BIASED* and *DEBIASED*, the same 200 SATIRICAL instances are selected, but only 100 of them are passed through the *debiasing pipeline*. Hence, the final *HYBRID* setup consists of 200 instances from NON-SATIRICAL corpora, 100 instances from original SATIRICAL corpora, and 100 instances from debiased SATIRICAL corpora.

After training, the models were evaluated in same-domain, cross-domain, and cross-lingual settings using the following test datasets:

Zaytung + AA: All instances from the *Turkish Satirical News Dataset* that are not used in

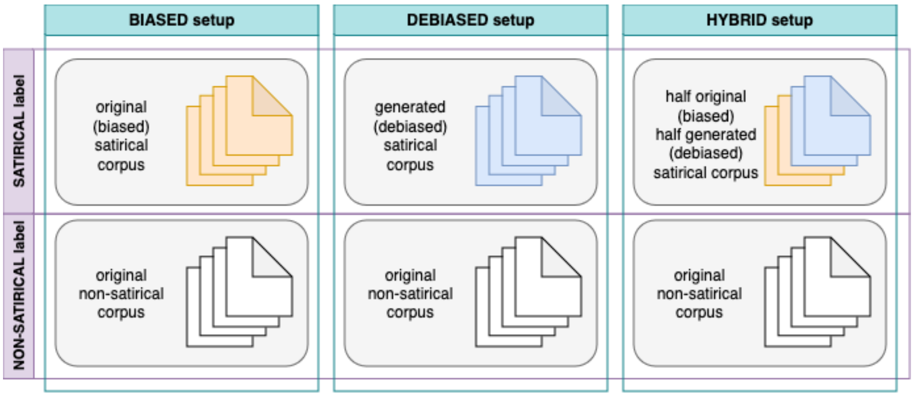

BIASED setup
DEBIASED setup
HYBRID setup
SATIRICAL label
NON-SATIRICAL label
original (biased) satirical corpus
generated (debiased) satirical corpus
half original (biased) half generated (debiased) satirical corpus
original non-satirical corpus
original non-satirical corpus
original non-satirical corpus

Figure 4: Training setups in the experiments

the training sets are included in this test set. Hence, this set consists of 4581 instances from the NON-SATIRICAL corpora and 2002 instances from the SATIRICAL corpora of the *Turkish Satirical News Dataset*.

The Onion + HuffPost: A fairly balanced set of 29000 English news article headlines from the American satirical news website the Onion[8] and HuffPost[9], taken from the openly available News Headlines Dataset For Sarcasm Detection[10].

IronyTR: IronyTR (Ozturk et al., 2021) is a Turkish social media irony detection dataset annotated by humans. It contains 300 Turkish ironic short texts and 300 non-ironic social media posts. This dataset is used to evaluate the cross-domain performance of the models trained under different settings.

5.3 Results

Table 4 presents the fine-tuning results of the selected language models on the *BIASED*, *DEBIASED*, and *HYBRID* setups which were evaluated on the Zaytung test set. According to the results, for each model except the Llama-3.1-8B, the *BIASED* setup achieved the highest F1-macro score. Since the training and test sets are from the same domain and the writing styles of the satirical and non-satirical news are significantly different, the outcome is expected. However, the proposed debiasing pipeline significantly reduced the F1-macro score across all language models. This reduction is attributed to the pipeline's goal of changing the writing style and diminishing the sarcastic tone in the texts, by making it more challenging for the models to differentiate between non-satirical and satirical news.

Secondly, Table 5 presents the cross-lingual evaluation results where the models trained on the Turkish dataset were tested using The Onion dataset. The results show that, except for Llama-3.1-8B, the proposed debiasing approach positively improved the F1-macro scores. In other words, the XLM-RoBERTa model achieved the highest score on the *HYBRID* dataset, which includes debiased instances, while the BERT models performed best on the *DEBIASED* setup.

Finally, Table 6 presents the models' performance in a cross-domain setting using the IronyTR dataset which contains ironic and non-ironic social media posts. Since the social media posts are short texts, whereas the training instances are long-form news articles, the models' performance was significantly lower than the results in Table 4. The proposed debiasing pipeline positively impacted the BERTurk and XLM-RoBERTa models; however, the highest F1-macro scores for multilingual BERT and Llama-3.1-8B were observed in the *BIASED* setup. While the Llama-3.1-8B model achieved its best F1-macro scores on both The Onion and IronyTR datasets using the *BIASED* setup, the scores were very close to those obtained in the other setups.

5.4 Discussion

Following the *DEBIASED* training, the masked language models demonstrated improved robustness in both cross-lingual (see Table 5) and cross-domain settings (see Table 6). However, for the Llama-3.1-8B model, *BIASED* training achieved the highest score, though the margin compared to other setups was minimal. More specifically in Table 5, *BIASED* setup outperformed *DEBIASED*

[8] https://theonion.com/

[9] https://www.huffpost.com/

[10] https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection/data

	BIASED				DEBIASED				HYBRID			
Models	**accuracy**	**precision**	**recall**	**f1-macro**	**accuracy**	**precision**	**recall**	**f1-macro**	**accuracy**	**precision**	**recall**	**f1-macro**
berturk	93.50%	93.95%	93.50%	**92.56%**	78.74%	77.90%	78.74%	72.26% (-20.30%)	90.52%	90.45%	90.52%	88.41% (-4.15%)
mbert-base	95.23%	95.59%	95.23%	**94.53%**	56.11%	56.21%	56.11%	56.01% (-38.55%)	94.68%	94.66%	94.68%	93.68% (-0.85%)
xlm-roberta large	97.83%	97.86%	97.83%	**97.39%**	93.61%	93.96%	93.61%	92.01% (-5.38%)	96.63%	96.73%	96.63%	95.90% (-1.49%)
llama-3.1-8B	67.26%	84.18%	67.26%	67.14%	65.28%	83.79%	65.28%	65.21% (-1.93%)	89.21%	91.40%	89.21%	**88.17%** (+21.03%)

Table 4: Evaluation on Zaytung + AA dataset

	BIASED				DEBIASED				HYBRID			
Models	**accuracy**	**precision**	**recall**	**f1-macro**	**accuracy**	**precision**	**recall**	**f1-macro**	**accuracy**	**precision**	**recall**	**f1-macro**
berturk	48.44%	48.09%	48.44%	47.78%	52.38%	52.68%	52.38%	**52.37%** (+4.59%)	50.41%	51.39%	50.41%	49.56% (+1.78%)
mbert-base	49.29%	48.44%	49.29%	47.34%	55.02%	56.26%	55.02%	**54.49%** (+7.15%)	56.08%	58.26%	56.08%	49.27% (+1.93%)
xlm-roberta large	52.45%	73.32%	52.45%	34.58%	56.53%	56.43%	56.53%	55.09% (+20.51%)	63.36%	63.72%	63.36%	**62.34%** (+27.76%)
llama-3.1-8B	70.95%	71.39%	70.95%	**70.70%**	69.54%	72.63%	69.54%	68.81% (-2.14%)	64.87%	66.36%	64.87%	63.23% (-7.72%)

Table 5: Evaluation on The Onion + HuffPost datasets

	BIASED				DEBIASED				HYBRID			
Models	**accuracy**	**precision**	**recall**	**f1-macro**	**accuracy**	**precision**	**recall**	**f1-macro**	**accuracy**	**precision**	**recall**	**f1-macro**
berturk	53.10%	53.28%	53.10%	52.67%	59.46%	60.81%	59.46%	**58.01%** (+5.34%)	51.59%	51.58%	51.59%	51.56% (-1.11%)
mbert-base	62.48%	64.51%	62.48%	**61.23%**	56.11%	56.21%	56.11%	56.01% (-5.22%)	58.29%	58.41%	58.29%	58.08% (-3.15%)
xlm-roberta large	52.93%	55.27%	52.93%	46.43%	64.99%	70.61%	64.99%	62.56% (+16.13%)	68.01%	68.77%	68.01%	**67.72%** (+21.29%)
llama-3.1-8B	64.58%	64.59%	64.58%	**64.55%**	62.58%	70.90%	62.58%	58.69% (-5.86%)	64.83%	64.98%	64.83%	64.47% (-0.08%)

Table 6: Evaluation on the IronyTR dataset

setup by 2.14%, while in Table 6, it outperformed the *HYBRID* setup by 0.08%. This performance in *BIASED* setups for LLama can be attributed to the model's pretraining knowledge. In other words, articles from The Onion, HuffPost, and instances of IronyTR might have been included in the pretraining data for the Llama model. Furthermore, Llama significantly outperformed the masked language models on The Onion + HuffPost dataset, further suggesting potential exposure during pretraining. Lastly, on *HYBRID* setup, XLM-RoBERTa outperformed the *BIASED* setup on both cross-lingual and cross-domain evaluations (see Table 5 and Table 6) with a significant margin. This result indicates that combining biased and debiased articles contributes to the model's robustness.

6 Conclusions

The problem of satire detection demands a human in the loop by its nature since the labeling process cannot be automated. The only automatization possible is finding a satirical resource (such as Zaytung for Turkish) and assuming all scraped content is satirical by default. Unfortunately, this causes the data to be biased stylistically and trickles down this bias to the model where the model learns to identify the style of the corpus instead of the satire.

This work proposes a debiasing method utilizing LLM-based text generation within ethical limits. We show that generating data that is stylistically neutral to replace the biased data in the training set decreases the model performance significantly and improves the cross-lingual and cross-domain robustness of the model for satire detection in Turkish. However, additional experimentation is needed to see if this method is generalizable as a debiasing method for different language tasks. Yet, the obtained results are promising to demonstrate the applicability of the proposed method.

7 Limitations

We tested a limited number of models which may not fully capture the variability across different models and configurations. Furthermore, there is a potential risk that some dataset instances may overlap with the training data of the LLMs (especially for the Llama-3.1-8B model) which could bias the evaluation results. Moreover, for the Zaytung dataset, the text field exceeded the sequence length of the masked language models (for BERT and RoBERTA). Therefore, we cropped the text fields for such instances.

It should also be noted that, using LLMs, specifically generative LLMs, ethical and environmental concerns should always be kept in mind. Generating textual data is an ethically convoluted topic, and should not be taken lightly. We believe that LLM-generated data should not be contextualized as if a real human has generated that content. These concerns may be limiting factors for the scalability of this study.

Acknowledgments

The authors would like to thank Mustafa Ozan Alpay and Selin Özgürsoy for their hard work and contributions to data collection and annotation.

This work is supported by the EuroHPC Development Access Call with Project ID DD-24-99. Parts of this research received the support of the EXA4MIND project, funded by the European Union's Horizon Europe Research and Innovation Programme, under Grant Agreement N° 101092944. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

Emre Can Acikgoz, Mete Erdogan, and Deniz Yuret. 2024. Bridging the bosphorus: Advancing turkish large language models through strategies for low-resource language adaptation and benchmarking. *arXiv preprint arXiv:2405.04685*.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Ulas Baran Baloglu, Bilal Alatas, and Harun Bingol. 2019. Assessment of supervised learning algorithms for irony detection in online social media. *1st International Informatics and Software Engineering Conference (UBMYK)*, pages 1–5.

Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. Modelling sarcasm in twitter, a novel approach. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49.

Ege Berk Buyukbas, Adnan Harun Dogan, Asli Umay Ozturk, and Pinar Karagoz. 2021. Explainability in irony detection. In *Big Data Analytics and Knowledge Discovery: 23rd International Conference, DaWaK 2021, Virtual Event, September 27–30, 2021, Proceedings 23*, pages 152–157. Springer.

Paula Carvalho, Bruno Martins, Hugo Rosa, Silvio Amir, Jorge Baptista, and Mário J Silva. 2020. Situational irony in farcical news headlines. In *International Conference on Computational Processing of the Portuguese Language*, pages 65–75. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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.

Gökhan Doğru, Adrià Martín Mor, and Anna Aguilar-Amat. 2018. Parallel corpora preparation for machine translation of low-resource languages: Turkish to english cardiology corpora. In *Proceedings of the LREC 2018 Workshop'MultilingualBIO: Multilingual Biomedical Text Processing'*, pages 12–15.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Paolo Rosso, and Véronique Moriceau. 2020. Irony detection in a multilingual context. In *European Conference on Information Retrieval*, pages 141–149. Springer.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41.

Yucheng Lin, Yuhan Xia, and Yunfei Long. 2024. Augmenting emotion features in irony detection with large language modeling. *arXiv preprint arXiv:2404.12291*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Aaron Maladry, Els Lefever, Cynthia Van Hee, and Veronique Hoste. 2023. A fine line between irony and sincerity: Identifying bias in transformer models for irony detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 315–324.

Aytuğ Onan and Mansur Alp Toçoğlu. 2020. Satire identification in turkish news articles based on ensemble of classifiers. *Turkish Journal of Electrical Engineering and Computer Sciences*, 28(2):1086–1106.

Reynier Ortega-Bueno, Francisco Rangel, D Hernández Farıas, Paolo Rosso, Manuel Montes-y-Gómez, and José E Medina Pagola. 2019. Overview of the task on irony detection in Spanish variants. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), Co-Located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS. Org*, volume 2421, pages 229–256.

Reynier Ortega-Bueno, Paolo Rosso, and Elisabetta Fersini. 2023. Cross-domain and cross-language irony detection: The impact of bias on models' generalization. In *International Conference on Applications of Natural Language to Information Systems*, pages 140–155. Springer.

Asli Umay Ozturk, Yesim Cemek, and Pinar Karagoz. 2021. Ironytr: Irony detection in turkish informal texts. *International Journal of Intelligent Information Technologies (IJIIT)*, 17(4):1–18.

Endang Wahyu Pamungkas and Viviana Patti. 2018. # NonDicevoSulSerio at SemEval-2018 task 3: Exploiting emojis and affective content for irony detection in english tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 649–654.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ipek Baris Schlicht, Defne Altiok, Maryanne Taouk, and Lucie Flek. 2024. Pitfalls of conversational llms on news debiasing. *arXiv preprint arXiv:2404.06488*.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts. *arXiv preprint arXiv:2401.13136*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405.

David Tomás, Reynier Ortega-Bueno, Guobiao Zhang, Paolo Rosso, and Rossano Schifanella. 2023. Transformer-based models for multimodal irony detection. *Journal of Ambient Intelligence and Humanized Computing*, 14(6):7399–7410.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2016a. Exploring the realization of irony in Twitter data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1794–1799.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2016b. Monday mornings are my fave:)# not exploring the automatic recognition of irony in english tweets. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2730–2739.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Junxin Liu, Zhigang Yuan, and Yongfeng Huang. 2018. THU-NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning. *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 51–56.

Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in Social Media: Definition, Manipulation, and Detection. *ACM SIGKDD Explorations Newsletter*, 21(2):80–90.

Rong Xiang, Xuefeng Gao, Yunfei Long, Anran Li, Emmanuele Chersoni, Qin Lu, and Chu-Ren Huang. 2020. Ciron: A new benchmark dataset for Chinese irony detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5714–5720. European Language Resources Association.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*.

Shiwei Zhang, Xiuzhen Zhang, Jeffrey Chan, and Paolo Rosso. 2019a. Irony detection via sentiment-based transfer learning. *Information Processing & Management*, 56(5):1633–1644.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Curating Human Annotations

As is, the satirical class of the *Turkish Satirical News Dataset* is labelled as satirical since it is known to be collected from a satirical online newspaper. However, it is not analyzed to see what properties of the news articles make them satirical in the first place.

To extend the usability fields of the curated dataset and obtain more information about the satirical corpus in the dataset as a baseline for explainability tasks, a subset of the SATIRICAL instances are further annotated by a human annotator. The annotation process is as follows:

1. The main annotator goes through the whole article body and identifies the REAL and FAKE parts.

2. The REAL and FAKE markings are done according to the objective facts and events. The annotator is asked to fact-check and cite related information as needed.

3. Four volunteers from different age demographics cross-check the annotations to have a higher coverage of news landscape knowledge.

4. News articles with annotations that have a unified agreement are accepted, the rest is discarded.

Finally, the human-annotated corpus consists of 40 satirical articles. Three selected annotations are shown in Figures 6a, 7a, and 5a. The red text stands for the FAKE parts of the article, whereas the blue parts are marked as REAL.

B Comparing Model Explanations with Human Annotations

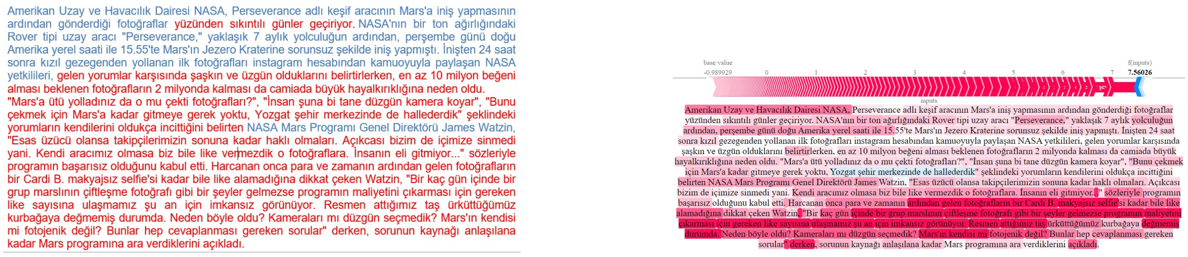
The curated *Turkish Satirical News Dataset* includes human annotations for 40 SATIRICAL articles for utilizing explainable AI and interpretable ML methods on trained models. However, to draw a comparison between the human annotations and model explanations, it is needed to define a relation between satirical and fake content. Considering the nature of satirical news articles, it is assumed that the parts that are labeled as FAKE in the annotation are likely to contribute to the satirical meaning of the article. This can be in the form of a fake person, a fake quote, or a fake event.

Similarly, the parts that are annotated to be REAL are less likely to contribute to the overall satire in the text. For example, the event described in an article may be real, therefore it can be annotated as REAL, but there may be a fake quote in the rest of the article that contributes to the satirical meaning. Following these parallels, a binary classifier is trained on *Turkish Satirical News Dataset*, and the model decisions are explained using SHAP.

The SHAP (Lundberg and Lee, 2017) explainability method uses Shapley values to understand the relative importance of different features for a prediction instance of a model. In other words, it assigns importance values to the features relative to each other that show their weight in the final decision.

As a binary classifier model to be explained using SHAP, BERTurk (Schweter, 2020) is fine-tuned. Later, SHAP-based explanations are extracted from the model. Three selected articles that have been classified correctly and have also been annotated by the human annotator are compared. One of these comparisons is reported in this section, and the other two are discussed in detail in Appendix B, with English translations for all three articles.

The selected explanation is shown in Figure 6b, and its human-annotated counterpart is shown in Figure 6a. The red highlights in the human annotation stand for the parts of the texts that are annotated as FAKE and the blue highlights specify the parts that are annotated as REAL. Similarly, for the SHAP output, red highlighted parts are explained as the *important* parts of the texts that the classifier focuses on when identifying a data instance as SATIRICAL. Blue highlights in the SHAP output indicate that those parts of the texts are pulling the label towards LEGITIMATE, and the parts that are



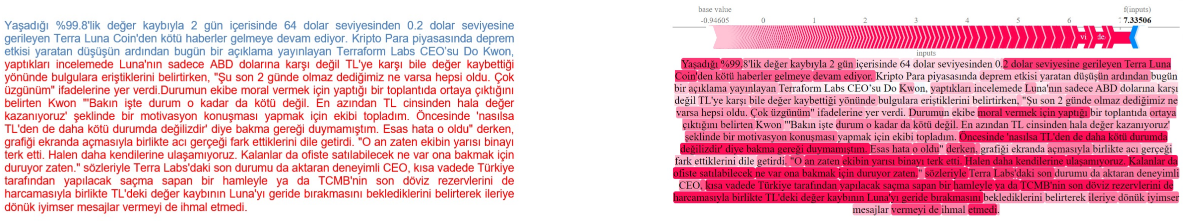
Amerikan Uzay ve Havacılık Dairesi NASA, Perseverance adlı keşif aracının Mars'a iniş yapmasının ardından gönderdiği fotoğraflar yüzünden sıkıntılı günler geçiriyor. NASA'nın bir ton ağırlığındaki Rover tipi uzay aracı "Perseverance," yaklaşık 7 aylık yolculuğun ardından, perşembe günü doğu Amerika yerel saati ile 15.55'te Mars'ın Jezero Kraterine sorunsuz şekilde iniş yapmıştı. İnişten 24 saat sonra kızıl gezegenden yollanan ilk fotoğrafları instagram hesabından kamuoyuyla paylaşan NASA yetkilileri, gelen yorumlar karşısında şaşkın ve üzgün olduklarını belirtirlerken, en az 10 milyon beğeni alması beklenen fotoğrafların 2 milyonda kalması da camiada büyük hayalkırıklığına neden oldu.

"Mars'a ütü yolladınız da o mu çekti fotoğrafları?", "İnsan şuna bi tane düzgün kamera koyar", "Bunu çekmek için Mars'a kadar gitmeye gerek yoktu, Yozgat şehir merkezinde de hallederdik" şeklindeki yorumların kendilerini oldukça incittiğini belirten NASA Mars Programı Genel Direktörü James Watzin, "Esas üzücü olansa takipçilerimizin sonuna kadar haklı olmaları. Açıkcası bizim de içimize sinmedi yani. Kendi aracımız olmasa biz bile like vermezdik o fotoğraflara. İnsanın eli gitmiyor..." sözleriyle programın başarısız olduğunu kabul etti. Harcanan onca para ve zamanın ardından gelen fotoğrafların bir Cardi B. makyajsız selfie'si kadar bile like alamadığına dikkat çeken Watzin, "Bir kaç gün içinde bir grup marslının çiftleşme fotoğrafı gibi bir şeyler gelmezse programın maliyetini çıkarması için gereken like sayısına ulaşmamız şu an için imkansız görünüyor. Resmen attığımız taş ürküttüğümüz kurbağaya değmemiş durumda. Neden böyle oldu? Kameraları mı düzgün seçmedik? Mars'ın kendisi mi fotojenik değil? Bunlar hep cevaplanması gereken sorular" derken, sorunun kaynağı anlaşılana kadar Mars programına ara verdiklerini açıkladı.

(a) Human annotation for the article

(b) SHAP output for the article

Figure 5: Human annotation and SHAP annotation for Sample Article (1)

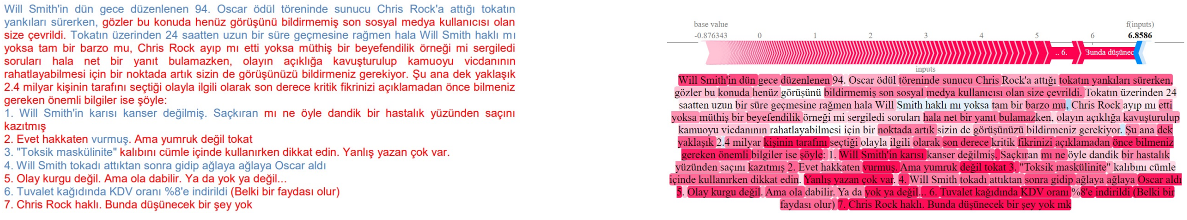


Yaşadığı %99.8'lik değer kaybıyla 2 gün içerisinde 64 dolar seviyesinden 0.2 dolar seviyesine gerileyen Terra Luna Coin'den kötü haberler gelmeye devam ediyor. Kripto Para piyasasında deprem etkisi yaratan düşüşün ardından bugün bir açıklama yayınlayan Terraform Labs CEO'su Do Kwon, yaptıkları incelemede Luna'nın sadece ABD dolarına karşı değil TL'ye karşı bile değer kaybettiği yönünde bulgulara eriştiklerini belirtirken, "Şu son 2 günde olmaz dediğimiz ne varsa hepsi oldu. Çok üzgünüm" ifadelerine yer verdi.Durumun ekibe moral vermek için yaptığı bir toplantıda ortaya çıktığını belirten Kwon "'Bakın işte durum o kadar da kötü değil. En azından TL cinsinden hala değer kazanıyoruz' şeklinde bir motivasyon konuşması yapmak için ekibi topladım. Öncesinde 'nasılsa TL'den de daha kötü durumda değilizdir' diye bakma gereği duymamıştım. Esas hata o oldu" derken, grafiği ekranda açmasıyla birlikte acı gerçeği fark ettiklerini dile getirdi. "O an zaten ekibin yarısı binayı terk etti. Halen daha kendilerine ulaşamıyoruz. Kalanlar da ofiste satılabilecek ne var ona bakmak için duruyor zaten." sözleriyle Terra Labs'daki son durumu da aktaran deneyimli CEO, kısa vadede Türkiye tarafından yapılacak saçma sapan bir hamleyle ya da TCMB'nin son döviz rezervlerini de harcamasıyla birlikte TL'deki değer kaybının Luna'yı geride bırakmasını beklediklerini belirterek ileriye dönük iyimser mesajlar vermeyi de ihmal etmedi.

(a) Human annotation for the article

(b) SHAP output for the article

Figure 6: Human annotation and SHAP annotation for Sample Article (2)



Will Smith'in dün gece düzenlenen 94. Oscar ödül töreninde sunucu Chris Rock'a attığı tokatın yankıları sürerken, gözler bu konuda henüz görüşünü bildirmemiş son sosyal medya kullanıcısı olan size çevrildi. Tokatın üzerinden 24 saatten uzun bir süre geçmesine rağmen hala Will Smith haklı mı yoksa tam bir barzo mu, Chris Rock ayıp mı etti yoksa müthiş bir beyefendilik örneği mi sergiledi soruları hala net bir yanıt bulamazken, olayın açıklığa kavuşturulup kamuoyu vicdanının rahatlayabilmesi için bir noktada artık sizin de görüşünüzü bildirmeniz gerekiyor. Şu ana dek yaklaşık 2.4 milyar kişinin tarafını seçtiği olayla ilgili olarak son derece kritik fikrinizi açıklamadan önce bilmeniz gereken önemli bilgiler ise şöyle:

1. Will Smith'in karısı kanser değilmiş. Saçkıran mı ne öyle dandik bir hastalık yüzünden saçını kazıtmış
2. Evet hakkaten vurmuş. Ama yumruk değil tokat
3. "Toksik maskülinite" kalıbını cümle içinde kullanırken dikkat edin. Yanlış yazan çok var.
4. Will Smith tokadı attıktan sonra gidip ağlaya ağlaya Oscar aldı
5. Olay kurgu değil. Ama ola dabilir. Ya da yok ya değil...
6. Tuvalet kağıdında KDV oranı %8'e indirildi (Belki bir faydası olur)
7. Chris Rock haklı. Bunda düşünecek bir şey yok

(a) Human annotation for the article

(b) SHAP output for the article

Figure 7: Human annotation and SHAP annotation for Sample Article (3)

not highlighted are not important for the decision of the model.

According to Figures 6a, 6b, 7a, 7b, 5a, and 5b it can be seen that the SHAP output and the human annotation overlap for most of the red highlights, meaning that the expected match between SATIRICAL and FAKE annotations is observed. On the other hand, this seems not to be the case for blue highlights, i.e. for the parts that are annotated as REAL by the human annotators. It is observed that the model sometimes considers these parts as an indication of the SATIRICAL label or does not use those parts in the prediction at all.

A closer look at article A (Figure 6a and Figure 6b) shows that the red highlights for both the human annotation and model explanation generally match, but the blue highlights of the human annotator, i.e. the parts that are annotated as real correspond to the parts that are highlighted as slightly red or neutral by the model. Ideally, we would expect neutrality or blue highlights in the corresponding parts of the SHAP output.

Additionally, a closer examination of article B (Figure 7a and Figure 7b) shows that the model and human annotator are in disagreement for most of the annotations. Even though the more saturated reds highlighted by the model match the red highlights of the annotator, that is not the case for all of the red highlights. The SHAP output even shows blue highlights where the corresponding human annotation is red, or vice versa.

Finally, a closer look at article C (Figure 5a and Figure 5b) shows that both the human annotations and the SHAP output are less continuous than the previous two comparisons. Here, the highlights mostly line up with the human annotations, but the model generally misses the sudden truth value changes in a sentence, which is expected.

Since the three examples all belong to the set of instances where the model predicts the class correctly, and the model already predicts almost every instance correctly, ideally it is expected to see a consistent overlap between the annotations and the SHAP explanations, which is not the case. This supports the claim that a model trained with a stylistically biased dataset may not be learning the representation of satire, but the style of the corpus instead.

Additionally, translations of Sample Articles (1) and (2) can be found in Appendix C. Translation for Sample Article (3) can be found in D.

C Prompts and Generated Articles in Turkish and English

Following two are the prompts in their original language, Turkish:

- **(Prompt 1)** *"Sana satirik bir haber vereceğim, adım adım bu haberdeki satirik unsurları kaldırmanı isteyeceğim. Önce bunun için haberden çıkarılması gereken cümleleri tespit et, sonra da cümleler çıkarılmış haliyle haberi tekrar yaz.*
  Haber metni:"

- **(Prompt 2)** *"Sana bir metin vereceğim, içindeki satirik cümleleri daha düz bir dile çevirip tekrar yaz.*
  Haber metni:"

C.1 Sample Article 1 generated with Prompt 1

In Turkish:

"Amerikan Uzay ve Havacılık Dairesi NASA, Perseverance adlı keşif aracının Mars'a iniş yapmasının ardından gönderdiği fotoğraflar yüzünden sıkıntılı günler geçiriyor. Mars programının başarıyla sonuçlanan inişinden sonra, NASA yetkilileri beklenen ilginin aksine, sosyal medyada düşük bir etkileşimle karşılaştı. Program direktörü James Watzin, beklenen ilgiyi alamamalarının hayal kırıklığına yol açtığını ve programın başarısız olduğunu kabul etti. Sorunun kaynağını anlamak için Mars programına ara verildi."

In English:

"The American Space and Aeronautics Agency NASA is going through difficult days due to the photos sent after the Perseverance rover successfully landed on Mars. Contrary to expectations, NASA officials encountered low engagement on social media following the successful landing of the Mars program. Program director James Watzin admitted that the lack of expected interest has led to disappointment and that the program has been deemed a failure. The Mars program has been paused to understand the source of the problem."

C.2 Sample Article 2 generated with Prompt 1 and Prompt 2

Original article in Turkish:

Yaşadığı %99.8'lik değer kaybıyla 2 gün içerisinde 64 dolar seviyesinden 0.2 dolar seviyesine gerileyen Terra Luna Coin'den kötü haberler gelmeye devam ediyor. Kripto Para piyasasında deprem etkisi yaratan düşüşün ardından bugün bir açıklama yayınlayan Terraform Labs CEO'su Do Kwon, yaptıkları incelemede Luna'nın sadece ABD dolarına karşı değil TL'ye karşı bile değer kaybettiği yönünde bulgulara eriştiklerini belirtirken, "Şu son 2 günde olmaz dediğimiz ne varsa hepsi oldu. Çok üzgünüm" ifadelerine yer verdi.

Durumun ekibe moral vermek için yaptığı bir toplantıda ortaya çıktığını belirten Kwon "'Bakın işte durum o kadar da kötü değil. En azından TL cinsinden hala değer kazanıyoruz' şeklinde bir motivasyon konuşması yapmak için ekibi topladım. Öncesinde 'nasılsa TL'den de daha kötü durumda değilizdir' diye bakma gereği duymamıştım. Esas hata o oldu" derken, grafiği ekranda açmasıyla birlikte acı gerçeği fark ettiklerini dile getirdi.

"O an zaten ekibin yarısı binayı terk etti. Halen daha kendilerine ulaşamıyoruz. Kalanlar da ofiste satılabilecek ne var ona bakmak için duruyor zaten." sözleriyle Terra Labs'daki son durumu da aktaran deneyimli CEO, kısa vadede Türkiye tarafından yapılacak saçma sapan bir hamleyle ya da TCMB'nin son döviz rezervlerini de harcamasıyla birlikte TL'deki değer kaybının Luna'yı geride bırakmasını beklediklerini belirterek ileriye dönük iyimser mesajlar vermeyi de ihmal etmedi.

Original article in English:

Following a 99.8% loss in value, where Terra Luna Coin dropped from $64 to $0.20 within two days, bad news continues to emerge. After this dramatic drop that shook the cryptocurrency market, Do Kwon, CEO of Terraform Labs, released a statement today. He noted that their investigation revealed Luna had lost value not only against the US dollar but also against the Turkish Lira. Kwon said, "Everything we thought couldn't happen in the last two days has happened. I am very sorry."

Kwon explained that the realization came during a meeting he held to boost the team's morale, saying, “I gathered the team to give a motivational speech along the lines of ‘Look, it's not so bad. At least we are still gaining value in TL terms.’ I hadn't felt the need to check if we were worse off compared to TL. That was the real mistake." He shared that the painful truth became apparent when they opened the graph on the screen.

"At that moment, half the team left the building. We still can't reach them. The remaining team members are just looking around the office to see if there's anything left to sell," Kwon described the current situation at Terra Labs. The experienced CEO also mentioned that they expect the depreciation of the TL to eventually surpass Luna, whether due to some absurd move by Turkey or the Central Bank of the Republic of Turkey depleting its remaining foreign currency reserves. He added that he has not neglected to give optimistic messages for the future.

C.2.1 Generated with Prompt 1

In Turkish:

"Terra Luna Coin son günlerde büyük bir değer kaybı yaşadı. Terraform Labs CEO'su Do Kwon, yaptıkları incelemede Luna'nın sadece ABD dolarına karşı değil TL'ye karşı da değer kaybettiğini belirtti. Bu durum, şirket çalışanları arasında moral bozukluğuna neden oldu. Kwon, kısa vadede TL'deki değer kaybının Luna'yı geride bırakmasını beklediklerini ifade etti."

In English:

"Terra Luna Coin has experienced a significant loss in value in recent days. Terraform Labs CEO Do Kwon stated that their investigation revealed that Luna has lost value not only against the US dollar but also against the Turkish Lira (TL). This situation has caused a morale decline among company employees. Kwon expressed that in the short term, they expect the depreciation of the TL to surpass that of Luna."

C.2.2 Generated with Prompt 2

In Turkish:

"Son iki günde %99.8 değer kaybı yaşayan Terra Luna Coin, 64 dolardan 0.2 dolara kadar geriledi ve kötü haberler gelmeye devam ediyor.
Bu büyük düşüşün ardından Terraform Labs CEO'su Do Kwon, bugün yaptığı açıklamada, Luna'nın sadece ABD dolarına değil, Türk Lirası'na karşı da değer kaybettiğini tespit ettiklerini söyledi. Kwon, "Bu son iki günde olmasını beklemediğimiz her şey oldu. Çok üzgünüm," dedi.
Kwon, durumu ekibe moral vermek için düzenlediği bir toplantıda fark ettiklerini belirtti. "Ekibe, 'En azından TL karşısında değer kazandık' demek istedim ama durumu kontrol etmemiştim. Asıl hata bu oldu," dedi. Grafiği açtıklarında ise acı gerçekle yüzleştiklerini ifade etti.
"Toplantı sırasında ekibin yarısı binayı terk etti ve hala onlara ulaşamıyoruz. Kalanlar ise ofiste satılacak bir şeyler arıyor," diye konuşan Kwon, kısa vadede TL'nin değer kaybının Luna'nınkini geride bırakabileceğini belirterek, geleceğe yönelik iyimser bir mesaj vermeyi de ihmal etmedi."

In English:

"In the past two days, Terra Luna Coin has experienced a 99.8% loss in value, dropping from \$64 to \$0.2, and bad news continues to emerge.
Following this major drop, Terraform Labs CEO Do Kwon stated in a recent announcement that they have found evidence that Luna has lost value not only against the US dollar but also against the Turkish Lira. Kwon said, "Everything we didn't expect to happen in these last two days happened. I'm very sorry."
Kwon mentioned that they realized the situation during a meeting he organized to boost the team's morale. "I wanted to tell the team, 'At least we're gaining value against the TL,' but I hadn't checked the situation. That was the real mistake," he said. He also mentioned that they faced the harsh reality when they opened the chart.
"During the meeting, half of the team left the building, and we still can't reach them. The rest are just looking for things to sell in the office," Kwon said. He also noted that in the short term, the loss of value of the TL might surpass that of Luna, and he did not forget to provide an optimistic message for the future."

C.3 Sample Article 4 generated with Prompt 2

Original article in Turkish:

Bu sabah Taraf gazetesi tarafından ortaya atılan "Tutsaklara karşılık Süleyman Şah Türbesi IŞİD'a verilecek" şeklindeki şok haber, Dışişleri Bakanlığı tarafından kesin bir dille yalanlandı. Önce internet sitesinden yapılan açıklama sonra da Basın Sözcüsü Kamuran Aydilen aracılığı ile kamuoyunu aydınlatan Dışişleri Bakanlığı, "Ortadaki yanlış anlaşılmaları gidermek için söylüyoruz, Süleyman Şah Türbesi'nin yıkılarak yerine AVM yapılması konusunda IŞİD'le görüştüğümüz doğru. Neticede türbe yıkımında kendilerinden daha tecrübeli bir ekip yok. Ancak bunun dışında herhangi bir pazarlık söz konusu değil" ifadeleri ile iddiaları reddetti.

Bakanlık binasında gazetecilerin sorularını yanıtlayan Bakanlık Sözcüsü Aydilen, türbenin yıkım ihalesi için IŞİD ile pazarlık masasında oturulduğunu itiraf ederken, konunun rehinlerle doğrudan bir ilgisi bulunmadığını ise şu sözlerle savundu:

"Arkadaşlar 12 yıllık iktidarımızda artık bizi biraz tanımış olmanız lazım. Bütün dünya bilir ki biz, öyle 49 kişi için bir karış toprak vermeyiz. Hele de öyle bir toprağı, tam kupon arazi orası, deli misiniz ya? Mümkün mü böyle bir pazarlık? Türbeyi de geç, sırf arsası 4 milyar dolar eder. Orada nöbet tutan askerlerimize de sorduk, çevrede başka AVM de yokmuş. 'Çarşı izninde gidecek yer bulamıyoruz' diyorlar. Şu inşaat bir başlasın, Allah'ın izniyle para basacak orası..."

IŞİD'ın özellikle türbe yıkım işinde uzmanlaşmış, işlerini severek yapan ve sahiplenen bir örgüt olduğunun altını çizen Basın Sözcüsü, "Şu an bizden haber bekliyorlar, tamam dediğiimiz anda havanlarla falan girişecekler. Alimallah 1 saatte taş üstüne taş koymayız dediler. Rehineler konusunu öyle özel olarak konuşmadık ama o konuda bir jest yaparlarsa biz bunu geri çevirmeyiz elbette. Neticede birlikte iş yapan insanlarız, yarın öbür gün başka yıkım ihaleleri de olur... Bunları da değerlendireceklerdir" ifadelerine yer verdi.

Mevcut anlaşmanın devletin kasasından bir kuruş çıkmadan halledileceğinin üzerinde duran Aydilen, yapılması planlanan AVM'nin detaylarını da basın mensuplarıyla paylaştı:

"Bakın buradan bööyle şimdiki türbenin kubbesi şeklinde bir tavan geliyor. Orası food court olacak... Alt katta SHAH'S SPORT adında bir fitness salonu ve atış poligonu var. Ta buraya kadar da meydan, forum mantığı gibi düşünün siz. Şimdi tabii aklınıza hemen ulaşım işi geliyor... Onu da düşündük. Hızlı treni 2017'de Marmaray'la Halkalı'ya bağladıktan sonra, Halkalı Ankara arası 4.5 saate inmiş olacak. Ankardan da ring seferiyle tak Halep'tesin. Son olarak Halep - Karakoza arası İDO'nun motorlarına binecek vatandaşlarımız anında AVM'de olacak. Bu kadar basit. Ayrıca oradaki askerlerimizi de özel güvenlik ve otopark görevlisi olarak AVM'de istihdam etmeyi düşünüyoruz. Gördüğünüz gibi bu projede kaybeden yok..."

Bir soru üzerine Suleymanium AVM'yi, yaşasaydı Suleyman Şah'ın da takdirle karşılayacağını sözlerine ekleyen Dışişleri Sözcüsü, son olarak şunları kaydetti:

"Yani düşünün tarihe geçmiş bir şahısınız, arkanızda bir tanecik kullanılmayan türbe kalıyor. Ne sineması var, ne otoparkı... Böyle mi anmalıyız ecdadımızı? Ayrıca son dönemde biliyorsunuz TOKİ'nin mevcut tarihi yapılar etrafında çeşitli çalışmaları mevcut. Sosyal medyada tarihi kümbetle iç içe geçmiş yurtlarımız büyük ilgi gördü. Bu şekilde alışveriş keyfini manevi iklimle birleştiren bir çalışma halkımızın da ilgisini çekecektir..."

Original article in English:

This morning, the shocking news reported by Taraf newspaper, claiming "The Süleyman Shah Tomb will be handed over to ISIS in exchange for the hostages," was firmly denied by the Ministry of Foreign Affairs. The Ministry, which clarified the issue first through an announcement on its website and later through its Spokesperson Kamuran Aydilen, stated: "To clear up the misunderstandings, we are saying this: it is true that we have discussed with ISIS the demolition of the Süleyman Shah Tomb and replacing it with a shopping mall. After all, there is no team more experienced than them when it comes to demolitions. However, there is no bargaining or negotiation beyond this."

While answering journalists' questions at the Ministry building, Spokesperson Aydilen admitted that negotiations with ISIS had taken place regarding the demolition of the tomb, but he defended that the issue had nothing to do with the hostages, saying:

"Friends, you should have known us by now after 12 years in power. The whole world knows that we wouldn't give up an inch of land for 49 people. Especially not such a prime piece of land, are you crazy? How could there be such a deal? Forget the tomb, just the land itself is worth 4 billion dollars. We even asked the soldiers stationed there, and they said there aren't any other shopping malls around. They say, 'We can't find a place to go on our leave.' Once construction starts, God willing, that place will be printing money..."

Emphasizing that ISIS is a group specialized in demolitions, especially of tombs, and one that enjoys and takes ownership of their work, the Spokesperson added: "Right now, they are waiting for our confirmation. As soon as we say 'yes,' they'll come in with mortars and the like. I swear, they said they'll finish the job in an hour. We didn't specifically discuss the hostages, but if they make a gesture in that regard, we certainly won't turn it down. After all, we're people who work together, and who knows, there may be other demolition tenders in the future… They'll certainly evaluate those as well."

Aydilen stressed that the current agreement would be handled without spending a penny from the state's coffers and shared details of the planned shopping mall with the press:

"Look, we're bringing in a ceiling shaped like the current dome of the tomb. That area will be a food court... On the lower floor, there will be a fitness center called SHAH'S SPORT and a shooting range. Think of it like a forum or square. Now, of course, you're probably thinking about transportation... We've thought of that too. After Marmaray connects to Halkalı in 2017, the travel time from Halkalı to Ankara will be reduced to 4.5 hours. From Ankara, a shuttle service will quickly take you to Aleppo. Lastly, our citizens traveling between Aleppo and Karakoza will take the IDO ferries and will be at the shopping mall in no time. It's that simple. We also plan to employ our soldiers stationed there as security and parking attendants in the shopping mall. As you can see, no one loses in this project..."

In response to a question, the Ministry Spokesperson added that the Süleymanium Shopping Mall would have been approved by Süleyman Shah himself, had he been alive:

"Imagine, you're a historical figure, and the only thing left behind is an unused tomb. It has no cinema, no parking lot... Is this how we should remember our ancestors? Also, as you know, TOKİ (Housing Development Administration of Turkey) has been working around existing historical structures. Our dormitories, integrated with the historical domes, have garnered a lot of attention on social media. This project, which combines the joy of shopping with a spiritual atmosphere, will also attract the public's interest..."

Generated article in Turkish:

"Taraf gazetesinin "Tutsaklara karşılık Süleyman Şah Türbesi IŞİD'e verilecek" iddiası Dışişleri Bakanlığı tarafından kesin bir dille yalanlandı. Bakanlık, Süleyman Şah Türbesi'nin yıkılarak yerine AVM yapılması konusunda IŞİD'le görüştüklerini, ancak bu görüşmenin rehinelerle ilgisi olmadığını belirtti.
Bakanlık Sözcüsü Kamuran Aydilen, "IŞİD yıkım işinde deneyimli bir ekip. Türbeyi yıkmak için onlarla görüşüyoruz ama toprak verme pazarlığı yok. Türbenin arsası çok değerli, rehineler için böyle bir pazarlık olmaz. Orada AVM yaparak ekonomiye katkı sağlamayı planlıyoruz" dedi.
Aydilen, yapılacak AVM'nin türbenin kubbesi şeklinde bir tavanı olacağını, food court, fitness salonu ve atış poligonu gibi bölümler içereceğini, ulaşımın hızlı tren ve İDO motorları ile sağlanacağını açıkladı. Askerlerin de AVM'de özel güvenlik ve otopark görevlisi olarak istihdam edileceğini ekledi. Ayrıca, bu projeyle kimsenin kaybetmeyeceğini ve halkın ilgisini çekeceğini belirtti."

Generated article in English:

"The claim made by Taraf newspaper that the Süleyman Shah Tomb will be handed over to ISIS in exchange for hostages has been strongly denied by the Ministry of Foreign Affairs. The ministry clarified that they are in talks with ISIS regarding the demolition of the tomb to build a shopping mall in its place, but emphasized that these discussions are unrelated to the hostages.
Ministry spokesperson Kamuran Aydilen stated, "ISIS has an experienced demolition team. We're negotiating with them to demolish the tomb, but there is no land-for-hostages deal. The land where the tomb is located is very valuable, and we wouldn't negotiate it for hostages. We plan to contribute to the economy by building a shopping mall there."
Aydilen also mentioned that the mall will have a dome-shaped ceiling inspired by the tomb's dome and will include sections such as a food court, fitness center, and shooting range. Transportation to the mall will be provided by high-speed trains and İDO ferries. He added that the soldiers stationed there will be employed as security personnel and parking attendants at the mall. Furthermore, he emphasized that this project would not cause any losses and would attract public interest."

D Other articles

D.1 Translation of Sample Article (3)

Warning: You're the Only One Who Hasn't Weighed In on Will Smith's Slap...

As the repercussions of Will Smith's slap on comedian Chris Rock at the 94th Oscars last night continue to ripple, all eyes have turned to you, the last social media user who has yet to express an opinion on the matter. Over 24 hours have passed since the slap, and the questions of whether Will Smith was justified or simply out of line, and whether Chris Rock was rude or demonstrated exemplary gentlemanliness, still lack clear answers. For the sake of clarifying the situation and easing the public conscience, it's time for you to share your view.

Before you disclose your crucial opinion on this event, which has involved approximately 2.4 billion people choosing sides, here are some important details you need to know:

1. Will Smith’s wife is not suffering from cancer. She shaved her head due to some trivial disease like alopecia. 2. Yes, he really did hit him. But it was a slap, not a punch. 3. Be careful when using the term "toxic masculinity" in a sentence. Many people spell it wrong. 4. After delivering the slap, Will Smith went on to cry and then won the Oscar. 5. The incident is not staged. But it could be. Or not... who knows. 6. The VAT rate on toilet paper has been reduced to 8% (Maybe this will help). 7. Chris Rock is right. There’s nothing more to think about.

BEIR-NL: Zero-shot Information Retrieval Benchmark for the Dutch Language

Nikolay Banar* **Ehsan Lotfi*** **Walter Daelemans**

CLiPS, University of Antwerp, Belgium
{nicolae.banari, ehsan.lotfi, walter.daelemans}@uantwerpen.be

Abstract

Zero-shot evaluation of information retrieval (IR) models is often performed using BEIR; a large and heterogeneous benchmark composed of multiple datasets, covering different retrieval tasks across various domains. Although BEIR has become a standard benchmark for the zero-shot setup, its exclusively English content reduces its utility for underrepresented languages in IR, including Dutch. To address this limitation and encourage the development of Dutch IR models, we introduce BEIR-NL by automatically translating the publicly accessible BEIR datasets into Dutch. Using BEIR-NL, we evaluated a wide range of multilingual dense ranking and reranking models, as well as the lexical BM25 method. Our experiments show that BM25 remains a competitive baseline, and is only outperformed by the larger dense models trained for retrieval. When combined with reranking models, BM25 achieves performance on par with the best dense ranking models. In addition, we explored the impact of translation on the data by back-translating a selection of datasets to English, and observed a performance drop for both dense and lexical methods, indicating the limitations of translation for creating benchmarks. BEIR-NL is publicly available on the Hugging Face hub[1].

1 Introduction

An increasing number of natural language processing (NLP) tasks require an information retrieval (IR) step to identify relevant pieces of text in a large corpus of documents. Therefore, IR models are crucial in various use cases, including question-answering (Chen et al., 2017), claim-verification (Thorne et al., 2018), and retrieval-augmented generation (Lewis et al., 2020).

Recently, IR has witnessed significant progress, driven mainly by advancements in large language models (LLMs; Zhao et al., 2024). Pre-trained on large corpora, these models can generate high-quality contextualized textual embeddings that capture semantic relationships beyond surface-level features like keywords. The produced vector representations demonstrate strong performance in IR tasks, as well as in other problems (Muennighoff et al., 2023) such as classification and clustering.

Benchmarking and evaluating such models is essential in sustaining advances in NLP research. Comprehensive benchmarks provide a standardized framework to assess the performance of models, identify their limitations, and guide the direction of future work. BEIR (Benchmarking IR; Thakur et al., 2021) was introduced to address this need in IR and became a standard benchmark in zero-shot evaluation, enabling the comparison of retrieval models in a unified framework. BEIR offers a diverse and heterogeneous collection of datasets covering various domains from biomedical and financial texts to general web content, and recently has been integrated into the broader MTEB benchmark (Massive Text Embedding Benchmark; Muennighoff et al., 2023), which measures the performance of textual embeddings on a broad range of tasks. While BEIR has substantially advanced the evaluation of IR models, its main limitation lies in the monolingual structure, which restricts its application for other languages.

In this work, we focus on extending the BEIR benchmark to Dutch, a resource-scarce language in IR research. By translating datasets from BEIR into Dutch, we aim to provide a foundation for evaluating IR models in this language. Our benchmark BEIR-NL facilitates zero-shot IR evaluation and supports the development of retrieval models tailored to Dutch. In addition, we conduct extensive evaluations of small and mid-range multilingual IR models, which support Dutch, including dense ranking and reranking models. We make the BEIR-NL benchmark available on the Hugging Face hub,

[*] indicates equal contribution

[1] https://huggingface.co/collections/clips/beir-nl-6756c81a8ebab4432d922a08

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 36–45
January 20, 2025. ©2025 Association for Computational Linguistics

ensuring that it inherits the same licenses as the datasets from BEIR (Appendix A).

2 Related Work

Recently, increasing efforts have been directed towards extending English or multilingual benchmarks to cover more languages. These efforts are primarily divided into two categories: (i) the existing (or to-be) human-annotated datasets are compiled into benchmarks, or (ii) existing benchmarks are automatically translated into new languages. The first approach provides high-quality datasets but requires substantial time and financial investment. The second approach is faster and more cost-effective, but the quality of translations can affect the overall quality of the benchmark and potentially lead to inaccurate model evaluations (Engländer et al., 2024). However, the recent availability of relatively cheap and high-quality machine translation solutions (thanks mainly to the LLM developments and advances) has made this an attractive and commercially feasible option, especially for large datasets and benchmarks. Below we outline relevant work focused on extending existing benchmarks to additional languages.

In generative benchmarking, Lai et al. (2023) utilized ChatGPT to translate three widely-used benchmark datasets for LLMs into 26 languages, to evaluate the performance of models for the Okapi framework. These datasets include ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and MMLU (Hendrycks et al.). Vanroy (2023) extended these datasets, along with TruthfulQA (Lin et al., 2022), to Dutch using ChatGPT. Subsequently, Thellmann et al. (2024) added GSM8K (Cobbe et al., 2021) to the mentioned benchmarking datasets and translated the entire collection into 21 European languages using DeepL.

Another branch of work focuses on extending MTEB (Muennighoff et al., 2023), which evaluates the quality of textual embeddings across multiple tasks. Xiao et al. (2023) extended this benchmark to Chinese (C-MTEB) by collecting 35 publicly-available Chinese datasets. MTEB-French (Ciancone et al., 2024) added 18 datasets in French to MTEB, including both original and DeepL-translated data. Building on MTEB, Wehrli et al. (2024) introduced six benchmarking datasets for clustering text embeddings in German. A Polish version, MTEB-PL (Poświata et al., 2024), consists of 28 datasets, with its retrieval part sourced from BEIR-PL (Wojtasik et al., 2024). ruMTEB (Snegirev et al., 2024) comprises 23 tasks in the MTEB format, with primarily original datasets in Russian, and with one translated using DeepL. SEB (Scandinavian Embedding Benchmark; Enevoldsen et al., 2024) represents 24 evaluation tasks for Scandinavian languages, incorporating a portion of existing translated datasets from MTEB.

Finally in IR, mMARCO (Bonifacio et al., 2021) extended the popular MSMARCO dataset (Bajaj et al., 2016) to multiple languages by translating queries and passages using Google Translate and Helsinki-NLP models (Tiedemann and Thottingal, 2020). Most related to our work, BEIR-PL (Wojtasik et al., 2024) translated a subset of the BEIR benchmark to Polish using Google Translate.

These efforts highlight the necessity of extending existing benchmarks to a multilingual context, enabling the evaluation of models across a wide range of languages. Building on the previous work, our study extends the BEIR benchmark to Dutch using machine translation, providing a valuable resource for evaluating IR models in this language.

3 Dataset

The original BEIR benchmark (Thakur et al., 2021) comprises 18 datasets, covering 9 different information retrieval tasks. Of these, 4 datasets are not publicly available, and therefore are removed from our selection for BEIR-NL. The remaining 14 datasets are listed in Table 1 along with their selected features and statistics. Since most retrieval models are trained on MSMARCO (Bajaj et al., 2016), we also report on its Dutch-translated version from mMARCO (Bonifacio et al., 2021), but do not include it for translation. We refer the reader to the BEIR paper (Thakur et al., 2021) for further descriptions and more details on each dataset.

3.1 Translation

The next step is translating the selected 14 datasets from English to Dutch. After considering commonly used options, we opted for Gemini-1.5-flash[2] which offers a good balance of speed, cost, and translation quality. We prompted the model to translate the inputs, providing it with the input type (query or document), and domain (4th column in Table 1) as context. We used the API in batch mode, which lowers the total cost to less than 450

[2] A small portion of translations were done using GPT-4o-mini and Google Translate, as Gemini declined to translate certain content and had occasional issues with tags in prompts.

Task	Dataset	Source	Domain	#Queries	#Docs	Avg. D/Q
Biomedical IR	TREC-COVID	Voorhees et al. (2021)	Biomedical	50	171K	493.5
	NFCorpus	Boteva et al. (2016)	Biomedical	323	3.63K	38.2
Question Answering	NQ	Kwiatkowski et al. (2019)	Wikipedia	3,452	2.68M	1.2
	HotpotQA	Yang et al. (2018)	Wikipedia	7,405	5.23M	2.0
	FiQA-2018	Maia et al. (2018)	Financial	648	57.6K	2.6
Argument Retrieval	ArguAna	Wachsmuth et al. (2018)	Miscellaneous	1,406	8.67K	1.0
	Touche-2020	Bondarenko et al. (2020)	Miscellaneous	49	383K	19.0
Duplicate-Question Retrieval	CQADupstack	Hoogeveen et al. (2015)	StackExchange	13,145	457K	1.4
	Quora	Thakur et al. (2021)	Quora	10,000	522K	1.6
Entity Retrieval	DBPedia	Hasibi et al. (2017)	Wikipedia	400	4.64M	38.2
Citation Prediction	SciDocs	Cohan et al. (2020)	Scientific	1,000	25.7K	4.9
Fact Checking	SciFact	Wadden et al. (2020)	Scientific	300	5.18K	1.1
	FEVER	Thorne et al. (2018)	Wikipedia	6,666	5.42M	1.2
	Climate-FEVER	Diggelmann et al. (2020)	Wikipedia	1,535	5.42M	3.0
Passage Retrieval	mMARCO	Bonifacio et al. (2021)	Miscellaneous	6,980	8.84M	1.1

Table 1: Statistics of datasets included in the BEIR-NL benchmark (plus mMARCO). The table highlights the number of queries and documents, as well as the average number of relevant documents per query (Avg. D/Q) (from Thakur et al. (2021)).

Euro. The exact prompts can be found in Appendix B.

To assess the translation quality, we randomly sampled 10 items from each dataset (140 in total) and asked a native Dutch speaker to check the translations against the original English text, and annotate instances for major (i.e. translation includes semantic addition or omission) or minor (i.e. translation is correct but too literal) issues. The results show major and minor issues in 2.2% and 14.8% of samples respectively, which means that almost 98% of the translated samples can be trusted for semantic accuracy. We will revisit this issue in the discussion section.

4 Experimental Setup

This section provides an overview of the experimental setup used to assess the performance of different models on BEIR-NL. We mostly follow the BEIR official repository[3] for zero-shot evaluation, using the provided code as much as possible but occasionally adapt it to specific requirements of the evaluated models. In the following, we describe the models, data processing steps, and evaluation metrics used in our experiments.

4.1 Models

We include models from three categories: lexical models, dense ranking models, and dense reranking models.

4.1.1 Lexical models

As the most popular lexical retrieval solution, BM25 (Robertson et al., 1994) relies on keyword matching and utilizes empirical word (or token) weighting schemes to determine the relevance of documents to a given query. Despite lexical gap issues, where the vocabulary used in queries can differ from that of relevant documents, BM25 remains a robust baseline for many retrieval tasks and was outperformed only recently by E5 (Wang et al., 2022) on the BEIR retrieval benchmark (Thakur et al., 2021) in zero-shot setting. Similarly to Wojtasik et al. (2024), we utilize the BM25 implementation from Elasticsearch for Dutch.

4.1.2 Dense ranking models

Dense ranking (or embedding) models encode an input sequence into a dense vector, which can be used to calculate similarity or relevance between sequences (query and document in our case). Inspired by recent related studies and the MTEB leaderboard[4], we select the following multilingual retrieval models for our zero-shot experiments[5]: mContriever (Izacard et al., 2022), LaBSE (Feng et al., 2022), LEALLA (Mao and Nakagawa, 2023), mE5 (Wang et al., 2024), BGE-M3 (Chen et al., 2024), DPR-XM (Louis et al., 2024), jina-embeddings-v3 (Sturua et al., 2024), and mGTE (Zhang et al., 2024). Table 2 lists these models along with a number of relevant features. Follow-

[3] https://github.com/beir-cellar/beir

[4] https://huggingface.co/spaces/mteb/leaderboard

[5] Due to computational limitations, we exclude larger models like e5-mistral-7b-instruct and bge-multilingual-gemma2.

Model	Based on	#Parameters	Dim	Max input	IR Finetuned
e5-multilingual-small	Multilingual-MiniLM	118M	384	512	Yes
e5-multilingual-base	XLMRoberta-base	278M	768	512	Yes
e5-multilingual-large	XLMRoberta-large	560M	1024	512	Yes
e5-multilingual-large-instruct	XLMRoberta-large	560M	1024	512	Yes
gte-multilingual-base	-	305M	768	8192	Yes
jina-embeddings-v3	XLMRoberta-large	572M	1024	8192	Yes
bge-m3	XLMRoberta-large	568M	1024	8192	Yes
dpr-xm	XMOD	852M (277M†)	768	512	Yes
LEALLA-small	LaBSE (distilled)	69M	128	512	No
LEALLA-base	LaBSE (distilled)	107M	192	512	No
LaBSE	-	471M	768	512	No
mContriever	Bert-multilingual-base	179M	768	512	No
bge-reranker-v2-m3	bge-m3	568M	1024	8192	Yes
jina-reranker-v2-base-multilingual	XLMRoberta-base	278M	768	1024	Yes
gte-multilingual-reranker-base	gte-multilingual-base	305M	768	8192	Yes

Table 2: Dense ranking (top) and reranking (bottom) models used in our experiments. ‘Dim’ is the dimension of the output embedding vector. LaBSE and gte-multilingual-base are trained from scratch. LEALLA is distilled from LaBSE, and the rest are fine-tuned from the model mentioned in the second column. †: dpr-xm is modular and uses 277M parameters during inference.

ing the convention, we do not impose any limits on the input length for these models, allowing them to handle truncation if necessary[6]. In all cases, cosine similarity is employed to score similarity between the normalized embeddings.

4.1.3 Zero-shot reranking models

Unlike ranking models that are employed in a bi-encoder setting, reranking models rely on cross-encoding the query and document, which can provide more accurate results at a higher computational cost. Consequently, reranking models are usually applied on the top outputs of a fast ranking model such as BM25.

We examine three popular multilingual reranking models, namely bge-reranker-v2-m3 (Chen et al., 2024), jina-reranker-v2-base-multilingual (Sturua et al., 2024), and gte-multilingual-reranker-base (Zhang et al., 2024) (see Table 2-bottom). Following the convention (Thakur et al., 2021), we apply these models on the top-100 documents retrieved by BM25, and evaluate the reranked output. We do not restrict the input length for the reranking models, leaving them to manage truncation.

4.2 Metrics

To assess the performance of our models, we employ two standard retrieval metrics: nDCG@10 and Recall@100. NDCG (normalized discounted cumulative gain) is a ranking-aware metric often used to report retrieval performance, especially on graded (non-binary) labels (Thakur et al., 2021). We also report recall, which, although ranking-agnostic, is a useful and relevant metric for practical settings like retrieval-augmented generation.

5 Results and Discussion

5.1 Retrieval Performance on BEIR-NL

Table 3 shows the retrieval performance of the selected models on the 14 subsets of BEIR-NL, in addition to MSMARCO. As mentioned before, MSMARCO is not part of our dataset, but considering its popularity in retrieval training, we include it in the evaluations (based on the Dutch-translated version from mMARCO (Bonifacio et al., 2021)).

The results show that BM25 still provides a competitive baseline, and in many cases is only outperformed by the larger dense models. The four recently released multilingual-e5-large-instruct, gte-multilingual-base, jina-embeddings-v3 and bge-m3 achieve the best overall performances, with multilingual-e5-large-instruct getting the highest Recall@100 on half of the datasets. We also observe a sizeable gap between the older ‘sentence embedding’ models, and the new generation of trained-for-retrieval models (see the last column in Table 2), with the latter achieving substantially higher results. However, based on their published metadata, the majority of these models have been at least partially exposed to BEIR datasets in their training process, which makes the comparison unfair (The corresponding *potentially inflated* results

[6] Considering the average document length in BEIR datasets, truncation is rarely needed for any of these models.

Model		MSMARCO	TREC-COVID	NFCorpus	NQ	HotpotQA	FiQA-2018	ArguAna	Touche-2020	CQADupstack	Quora	DBPedia	SciDocs	SciFact	FEVER	Climate-FEVER
BM25	NDCG@10	16.87	63.37	30.54	25.09	53.62	18.73	41.76	28.15	27.77	65.92	25.46	11.44	61.13	60.65	12.09
multilingual-e5-small		30.85†	41.74	24.10	27.03†	53.30†	20.39	44.76	16.04	28.51	79.85†	25.89	6.58	58.82	56.69†	14.08
multilingual-e5-base		32.79†	40.68	24.17	36.06†	60.87†	23.76	47.06	10.29	30.36	81.02†	28.74	10.53	67.23	58.52†	16.31
multilingual-e5-large		**37.51**†	69.72	28.06	49.15†	67.95†	31.84	48.90	22.18	31.92	82.01†	**38.67**	11.95	68.38	72.73†	13.76
multilingual-e5-large-instruct		34.35†	71.22	31.08	**55.79**†	65.97†	**37.93**	50.32	26.67	36.95	83.54†	38.24	**18.07**	69.10	79.39†	21.05
gte-multilingual-base		27.19†	53.36	27.97†	47.42†	58.53†	29.45	**52.85**†	22.60 †	31.59†	81.25†	36.46†	15.86	64.41	82.68†	17.53
jina-embeddings-v3		26.05†	54.46	29.84	37.26†	51.82	35.71	52.23	15.05	36.16	82.92	30.71	18.42	64.90	68.88	19.54
bge-m3		31.96†	48.22	27.90	51.92†	65.20†	32.60	52.16	22.68	34.75	**83.72**	35.46	14.41	62.83	76.08	**26.39**
dpr-xm		28.46†	40.86	18.58	28.56	26.34	13.98	26.91	15.99	18.73	74.70	21.07	8.64	34.29	49.46	11.16
LEALLA-small		3.95	13.32	5.56	5.11	12.18	3.41	19.25	5.65	13.14	68.50	9.60	3.70	12.98	7.08	0.34
LEALLA-base		5.60	14.44	6.09	7.77	17.46	3.75	24.97	5.00	14.34	70.87	13.40	3.09	7.13	7.46	1.15
LaBSE		6.87	18.50	13.54	11.24	18.64	7.38	39.15	4.67	19.66	75.55	15.27	6.32	39.07	12.51	3.85
mContriever		7.46†	17.51	13.36	10.50	27.84	5.41	39.60	6.15	12.81	72.90	15.58	4.93	37.89	21.51	3.08
BM25 + bge-reranker		31.80†	76.47	**33.78**	51.28†	**71.78**†	30.41	47.27	**33.78**	31.70	76.81	37.84	13.88	69.94	84.17	25.60
BM25 + jina-reranker		31.93†	**76.83**	33.19	49.07†	70.57	30.86	48.53	30.96	34.06	79.44	36.26	14.49	**70.68**	**85.17**	22.56
BM25 + gte-reranker		28.90†	76.24	28.26†	47.85†	70.43†	24.13	46.74†	28.26†	25.69†	74.95†	36.67†	13.22	68.37	85.13†	22.96
BM25	Recall@100	51.20	10.52	22.16	65.57	70.54	42.83	92.32	44.16	54.77	88.66	36.92	26.49	83.42	89.20	30.42
multilingual-e5-small		74.63†	7.89	23.56	60.70†	69.45†	47.10	94.59	38.18	56.99	97.51†	35.83	22.93	87.67	85.83†	40.47
multilingual-e5-base		77.39†	6.58	22.09	73.61†	76.24†	55.02	95.59	32.96	60.65	97.93†	39.40	29.78	91.00	89.98†	42.69
multilingual-e5-large		**82.71**†	13.31	27.34	83.49†	**82.21**†	61.81	96.37	43.65	63.30	98.66†	47.26	30.42	92.27	93.08†	32.68
multilingual-e5-large-instruct		80.89†	**14.48**	**28.88**	**92.39**†	80.55†	68.70	98.86	46.97	70.56	98.83†	**49.66**	40.80	**93.67**	**94.53**†	**46.05**
gte-multilingual-base		70.29†	10.74	27.89†	85.39†	70.08†	61.53	97.87†	41.12†	66.14†	98.12†	44.11†	37.43	91.00	94.32†	40.40
jina-embeddings-v3		73.43†	11.74	26.50	84.43†	68.04	**69.98**	**98.93**	37.69	**72.62**	98.58	42.22	**42.64**	91.17	93.04	44.98
bge-m3		77.71†	9.43	25.20	89.62†	80.20†	63.41	97.44	**48.70**	66.89	**98.85**	46.30	35.02	91.93	94.11	56.54
dpr-xm		67.77†	5.78	17.95	62.42	38.31	33.81	78.73	36.46	41.94	93.41	22.25	19.36	67.26	76.17	28.54
LEALLA-small		15.99	1.44	9.12	19.99	23.48	12.19	56.47	9.89	32.58	91.41	13.38	12.62	42.81	14.79	1.81
LEALLA-base		22.12	1.61	9.92	27.45	30.32	13.04	61.30	8.39	33.89	93.13	18.80	10.73	34.18	14.97	2.61
LaBSE		26.71	1.97	16.05	41.68	33.56	25.57	87.98	10.09	47.06	95.87	22.91	21.50	74.67	36.48	15.24
mContriever		32.06†	1.71	16.81	40.42	45.97	20.36	91.61	12.06	35.91	94.48	25.25	18.56	74.24	48.31	10.29

Table 3: Performance of selected models on the BEIR-NL benchmark (plus MSMARCO), measured by NDCG@10 (top) and Recall@100 (bottom).† indicates results that are (or are highly likely to be) inflated because of potential contamination of the model with in-domain data for a given dataset, based on available descriptions from the corresponding work (i.e. they are highly unlikely to be zero-shot). bge-reranker, jina-reranker, and gte-reranker refer to bge-reranker-v2-m3, jina-reranker-v2-base-multilingual, and gte-multilingual-reranker models, respectively.

are marked with a † in the table.). In other words, in these cases the evaluation could not be considered proper zero-shot.

Finally, the last three rows of the top section in Table 3 (NDCG@10 results) show the performance of the reranking models when used in combination with BM25 as the first-step ranker. As demonstrated, this approach can often offer a competitive edge over the best ranking models.

5.2 Comparison with BEIR and BEIR-PL

Since BEIR-NL is a translated benchmark, we can compare the performance of the retrieval methods on parallel subsets in different languages, including the (translated) Polish version, BEIR-PL (Wojtasik et al., 2024).

Tables 4 and 5 show this comparison for BM25 and gte-multilingual-base, across the subsets for which performance data is publicly available[7]. As Table 4 reveals, BM25 performs comparably on BEIR-NL and BEIR-PL subsets, with a marginal overall advantage for BEIR-NL. However, these numbers lag behind the BM25 performance on the original BEIR dataset by 6-7 points in NDCG@10 and Recall@100. One potential reason for this drop is the lexical mismatch between the translated query and relevant passages since queries and passages are translated independently[8] (Bonifacio et al., 2021). Table 5 shows that the performance difference persists with dense models (e.g. gte-multilingual-base). Here, the discrepancy can be attributed to both the data (translation quality) and model (higher competence in English compared to other languages).

5.3 Impact of Translation

To isolate the semantic effect of translation (from that of the model/language) we back-translate a subset of 5 BEIR-NL datasets to English using the same translation pipeline, and compare the performance of lexical and dense models on this version against the original one. Table 6 shows the results (NDCG@10), which indicate an average drop of 1.9 and 2.6 points for the lexical (BM25) and dense model (gte-multilingual-base) respectively. Since the model-language competence factor is absent here, this drop can be considered a proxy for the impact of translation on the benchmark quality and/or reliability.

6 Conclusions and Future Work

In this work, we introduced BEIR-NL, an automatically translated version of the BEIR benchmark into Dutch, which aims to address the need for the evaluation of IR models in this language. Using BEIR-NL, we conducted extensive zero-shot evaluations for various models, including one lexical model as well as small and mid-range dense retrieval and reranking models. These experiments showed that larger dense IR models generally outperform BM25, while BM25 remains a competitive baseline for smaller models. Furthermore, combining BM25 with reranking models results in performance comparable to the best dense retrieval models.

We also observed several challenges, including the impact of translation on retrieval performance and the risk of in-domain data contamination in IR models. These issues might affect the reliability of zero-shot evaluations on this benchmark and highlight the need for creating native Dutch resources, which we leave for future work.

BEIR-NL fills a critical gap in the evaluation of Dutch IR models and sets a foundation for further development of IR benchmarks in Dutch. By making BEIR-NL publicly available, we aim to support future research and encourage the development of retrieval models for this language.

Limitations

Besides the issues originated from translation (which we briefly addressed before), here we discuss other important limitations pertinent to this work.

Native Dutch Resources. While BEIR-NL provides a benchmark for evaluating IR models in Dutch, it relies on translations from the original BEIR, which is exclusively in English. This lack of native Dutch datasets limits the ability of BEIR-NL to fully represent and reflect the linguistic nuances and cultural context of the language, and therefore the complexities of Dutch IR, especially in domain-specific contexts with local terminology and knowledge.

Data Contamination. Many modern IR models are trained on massive corpora that might include content from BEIR. Table 3 indicates multiple models that have (or might have) been exposed to in-domain contamination for a given dataset. This can result in inflated performance –as models might have already seen the relevant data during

[7] BEIR-PL only covers 10 of the 14 public BEIR datasets.

[8] Assuming a uniform BM25 performance for different languages, which is not trivial.

Metric	Benchmark	TREC-COVID	NFCorpus	NQ	HotpotQA	FiQA-2018	ArguAna	CQADupstack	DBPedia	SciDocs	SciFact	Average
NDCG@10	BEIR-NL	63.4	30.5	25.1	53.6	18.7	41.8	27.8	25.6	11.4	61.1	35.9
	BEIR-PL	61.0	31.9	20.1	49.2	19.0	41.4	28.4	22.9	14.1	62.5	35.1
	BEIR (EN)	68.9	34.3	32.6	60.2	25.4	47.2	32.5	32.1	16.5	69.1	41.9
Recall@100	BEIR-NL	10.5	22.2	65.6	70.5	42.8	92.3	54.8	36.9	26.5	83.4	50.6
	BEIR-PL	10.1	24.6	57.9	67.1	44.1	93.5	53.9	30.1	33.0	88.4	50.3
	BEIR (EN)	11.7	26.0	78.3	76.3	54.9	95.2	62.1	43.5	36.8	92.0	57.7

Table 4: BM25 performance on the overlapping subset of BEIR-NL, BEIR-PL, and original BEIR, for which performance data is publicly available. Results for BEIR-PL and BEIR are from Wojtasik et al. (2024).

Metric	Benchmark	TREC-COVID	NFCorpus	NQ	HotpotQA	FiQA-2018	ArguAna	DBPedia	SciDocs	SciFact	Average
NDCG@10	BEIR-NL	53.4	28.0	47.4	58.5	29.4	52.9	36.5	15.9	64.4	42.9
	BEIR-PL	59.4	26.8	43.1	56.9	29.0	53.2	32.5	14.2	58.9	41.6
	BEIR (EN)	57.6	36.6	58.1	63.0	45.0	58.2	40.1	18.2	73.4	50.0

Table 5: Performance of gte-multilingual-base on the overlapping subset of BEIR-NL, BEIR-PL, and original BEIR, for which performance data is publicly available. Results for BEIR-PL and BEIR are sourced from the MTEB leaderboard.

Model	BEIR	NFCorpus	FiQA-2018	ArguAna	SciDocs	SciFact	Average	Δ_{tr}
BM25	original	34.3	25.4	47.2	16.5	69.1	38.5	-
	back-translated	32.4	22.0	45.2	15.1	68.2	36.6	-1.9
gte-multilingual-base	original	36.7	45.0	58.2	18.2	73.4	46.3	-
	back-translated	32.6	40.7	55.0	18.3	71.7	43.7	-2.6

Table 6: NDCG@10 results for BM25 and gte-multilingual-base on selected datasets from the original BEIR, and their back-translated version (from Dutch to English). Δ_{tr} is the change in average performance due to back translation.

different phases of training– raising concerns about the validity of zero-shot evaluations. Ensuring a truly zero-shot evaluation is a difficult challenge, as many IR models lack transparency regarding the exact composition of training corpora.

Benchmark Validity Over Time. BEIR has become a standard benchmark to evaluate the performance of IR models, attracting a large number of evaluations over time. This extensive usage introduces the risk of overfitting, as researchers might unintentionally train models tailored to perform well on BEIR rather than on broader IR tasks. In addition, advances in IR models and evaluation needs might outpace the benchmark, making it less representative and less relevant. As a result, the relevance and validity of BEIR as well as BEIR-NL may diminish over time.

Acknowledgments

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. We would like to thank Jens Van Nooten for assessing the quality of the translations. In addition, we acknowledge the use of the GPT-4o model for assisting with error checking and proofreading of this paper.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, et al. 2020. Overview of touché 2020: argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 384–395. Springer.

Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, , Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of ms marco passage ranking dataset. *Preprint*, arXiv:2108.13897.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 716–722. Springer.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Mteb-french: Resources for french sentence embedding evaluation and analysis. *arXiv preprint arXiv:2405.20468*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. *arXiv preprint arXiv:2406.02396*.

Leon Engländer, Hannah Sterz, Clifton Poth, Jonas Pfeiffer, Ilia Kuznetsov, and Iryna Gurevych. 2024. M2qa: Multi-domain multilingual question answering. *arXiv preprint arXiv:2407.01091*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: a test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1265–1268.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Doris Hoogeveen, Karin M Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–8.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuật Nguyễn, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Antoine Louis, Vageesh Saxena, Gijs van Dijck, and Gerasimos Spanakis. 2024. Colbert-xm: A modular multi-vector representation model for zero-shot multilingual information retrieval. *arXiv preprint arXiv:2402.15059*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Zhuoyuan Mao and Tetsuji Nakagawa. 2023. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1886–1894, Dubrovnik, Croatia. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. Pl-mteb: Polish massive text embedding benchmark. *arXiv preprint arXiv:2405.10138*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov. 2024. The russian-focused embedders' exploration: rumteb benchmark and russian embedding model design. *arXiv preprint arXiv:2408.12503*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards multilingual llm evaluation for european languages. *arXiv preprint arXiv:2410.08928*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. 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 809–819.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd annual conference of the European Association for Machine Translation*, pages 479–480.

Bram Vanroy. 2023. Language resources for dutch large language modelling.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Silvan Wehrli, Bert Arnrich, and Christopher Irrgang. 2024. German text embedding clustering benchmark. *arXiv preprint arXiv:2401.02709*.

Konrad Wojtasik, Kacper Wołowiec, Vadim Shishkin, Arkadiusz Janz, and Maciej Piasecki. 2024. Beir-pl: Zero shot information retrieval benchmark for the polish language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2149–2160.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *Preprint*, arXiv:2407.19669.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

A Appendix: Licenses

The BEIR repository on Hugging Face[9] reports that the following datasets are distributed under the CC BY-SA 4.0 license: NFCorpus, FiQA-2018, Quora, Climate-Fever, FEVER, NQ, DBPedia, ArguAna, Touché-2020, SciFact, SCIDOCS, HotpotQA, TREC-COVID. The only one exception is CQADupStack[10] with the Apache License 2.0 license.

B Appendix: Translation Prompts

We prompt Gemini-1.5-flash with the following instructions (temperature = 0).

Query Prompt:"Translate to English the QUERY from the {domain} domain. Provide only the translation. QUERY:\n ['{query}']".

Document Prompt:"Translate to English the DOCUMENT from the {domain} domain. Provide only the translation. DOCUMENT:\n ['<title> {title} <title\> <body> {document} <body\>']".

[9] https://huggingface.co/datasets/BeIR/

[10] https://github.com/D1Doris/CQADupStack

Refining Dimensions for Improving Clustering-based Cross-lingual Topic Models

Chia-Hsuan Chang[1], Tien-Yuan Huang[2], Yi-Hang Tsai[2], Chia-Ming Chang[2], San-Yih Hwang[2]
[1]Department of Biomedical Informatics & Data Science, Yale University
New Haven, CT 06510, United States
[2]Department of Information Management, National Sun Yat-sen University
Kaohsiung 80424, Taiwan
Correspondence: shane.chang.tw@gmail.com, syhwang@mis.nsysu.edu.tw

Abstract

Recent works in clustering-based topic models perform well in monolingual topic identification by introducing a pipeline to cluster the contextualized representations. However, the pipeline is suboptimal in identifying topics across languages due to the presence of language-dependent dimensions (LDDs) generated by multilingual language models. To address this issue, we introduce a novel, SVD-based dimension refinement component into the pipeline of the clustering-based topic model. This component effectively neutralizes the negative impact of LDDs, enabling the model to accurately identify topics across languages. Our experiments on three datasets demonstrate that the updated pipeline with the dimension refinement component generally outperforms other state-of-the-art cross-lingual topic models [1].

1 Introduction

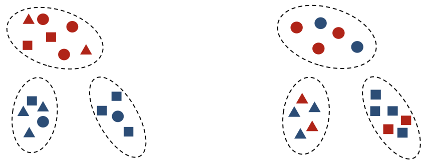
Traditional cross-lingual topic models (CLTM) rely on additional resources to identify topics across languages. Based on the types of resources, CLTMs can be categorized into document and vocabulary-linking models. The document-linking models require parallel or comparable corpora to model the co-occurring word statistics across languages and infer the cross-lingual topics (Mimno et al., 2009; Piccardi and West, 2021). The vocabulary-linking models are more resource-efficient than their document-linking counterpart because they only require a bilingual dictionary (i.e., a set of translation entries). However, vocabulary-linking models often result in monolingual topics (Hu et al., 2014; Hao and Paul, 2020; Wu et al., 2023) when the dictionary is of limited coverage to the target corpus. Several studies proposed to link word embedding spaces across languages to decrease the effort of compiling a well-covered dictionary. When the assumption of shared structures across spaces (i.e., isomorphism) holds, a small number of translation entries will be sufficient to identify topics across languages (Chang et al., 2018; Yuan et al., 2018; Chang and Hwang, 2021). However, the word spaces of different languages seldom share the same structure in practice, especially for languages that are distantly related, and iterative human involvement is still required for acquiring a quality dictionary.

The recent development of multilingual language models (MLM), e.g., mBERT, XLM-R, and GPT models, attracts attention from the natural language processing community. MLM learns the language-agnostic representations without any additional resources (Pires et al., 2019a; Dufter and Schütze, 2020), which has the potential to realize the zero-shot topic identification across languages (Bianchi et al., 2021), thereby reducing efforts on data preparation. Recent studies increasingly favor the clustering-based topic model due to its superior performance and higher efficiency (Sia et al., 2020; Grootendorst, 2022; Zhang et al., 2022). The clustering-based topic model adopts a pipeline (see Sec. 2.1) to leverage the induced representations of language models for topic identification. MLMs can be directly applied to the pipeline of clustering-based topic modeling for cross-lingual topic identification. However, the current pipeline is hindered by the existence of language-dependent dimensions (LDDs) in the representations generated by MLMs, which makes the representations sensitive to languages and hinders the pipeline from identifying topics across languages. As depicted in Fig. 1a, the current pipeline with MLM tends to cluster documents by languages rather than semantic meanings. We also report the qualitative result of misaligned topics generated using BERTopic (Grootendorst, 2022), an accessible implementation for clustering-based topic modeling, in Table 1. Ideally, topic clusters should group

[1] Our code and data are available at `https://github.com/Text-Analytics-and-Retrieval/Clustering-based-Cross-Lingual-Topic-Model`.

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 46–56
January 20, 2025. ©2025 Association for Computational Linguistics

documents based on their semantic meanings, as illustrated in Fig. 1b.



(a) Topic clusters grouped by languages (b) Topic clusters grouped by semantic meanings

Figure 1: Two resultant scenarios of clustering-based topic model. Different shapes indicate the documents discussing various topics, while different colors represent documents of different languages.

To mitigate such a problem, this study proposes adding a new dimension refinement component into the pipeline to neutralize the impacts of LDDs from the representations. Specifically, we utilize singular value decomposition (SVD) to identify the LDDs and offer two implementations of the dimension refining component: unscaled SVD (u-SVD) and SVD with language dimension removal (SVD-LR). The contributions of this study are threefold:

1. We observe and identify the negative impacts of LDDs on the pipeline of the clustering-based topic model in a cross-lingual topic identification task.

2. We introduce a dimension refinement component, implemented by either u-SVD or SVD-LR, into the current pipeline of the clustering-based topic model, which enables it to identify topics across languages.

3. Our updated pipeline of the clustering-based topic model is shown to outperform the other state-of-the-art CLTMs on three datasets.

2 Methodology

2.1 Background: Pipeline of Clustering-based Topic Model

The pipeline of clustering-based topic model (Grootendorst, 2022; Zhang et al., 2022) contains four steps: Document Embedding Generation → Dimension Reduction → Document Clustering → Cluster Summarization. The first step adopts a pre-trained language model to embed documents into contextualized representations. The next step, Dimension Reduction, reduces the dimension of the representations for speeding up the subsequent clustering process. The Document Clustering step applies some clustering techniques, e.g., K-Means (Zhang et al., 2022), to the reduced representations for topic cluster identification. The last step, Cluster Summarization, reconstructs topic-word distribution by using word importance ranking metric, e.g., c-TF-IDF (Grootendorst, 2022), on each topic cluster. c-TF-IDF calculate the importance of the word w in the cluster k by

$$\mathrm{tf}_{w,k} \times \log(1 + \frac{A}{f_w}), \qquad (1)$$

where $\mathrm{tf}_{w,k}$ is the word frequency of w in the document cluster k, A is the average word frequency of all clusters, and f_w is the frequency of word w across clusters. The higher value means the word w is more representative to a cluster k.

2.2 Pipeline Adaption for Cross-lingual Topic Identification

To adapt the current pipeline for cross-lingual topic identification, MLMs, such as Distilled XLM-R (Reimers and Gurevych, 2020; Conneau et al., 2020) and Cohere multilingual model, can be used in step 1 for embedding documents into language-agnostic representations $E \in R^{m \times d}$, where m is number of documents and d is dimension of representations. However, we observe that a number of dimensions of MLMs' representations retain language information. These dimensions are denoted as language-dependent dimensions (LDDs). To illustrate, we group documents written in language $l \in \{l_1, l_2\}$ and look into their representations. Let $e_i^l \in R^{m^l \times 1}$ be the values of i'th dimension for m^l documents written in l. We compare the values of each dimension $i \in d$ across two languages l_1 and l_2 by performing a two-sample t-test on $e_i^{l_1}$ and $e_i^{l_2}$. We then sort all dimensions based on the corresponding t-statistics in descending order. As the larger t-statistic indicates the larger mean value difference across languages, we hereby identify LDDs. As shown in the upper-left subplot of Fig. 2, the original MLM embeddings show notable distinctions for documents written in two different languages, suggesting the presence of LDDs within the original embeddings. Furthermore, after applying UMAP, a dimension reduction approach used by previous cluster-based topic models (Grootendorst, 2022; Zhang et al., 2022), even more significant LDDs are present (see the upper-right subplot of Fig. 2). This is likely to occur as UMAP focuses on capturing the local structure (McInnes et al., 2020).

Table 1: Top representative words of five sampled topics generated from BERTopic (Grootendorst, 2022) with default parameters. We first use Cohere multilingual model to embed the Airiti dataset (Chang et al., 2020) and then employ BERTopic to generate topics.

Topic#1	cell, protein, expression, induce, gene, mouse, find, show, study, treatment

Topic#2	細胞(cell), 蛋白(protein), 表現(expression), 基因(gene), 抑制(inhibition)
	蛋白質(protein), 我們(we), 發現(discover), 調控(control), 病毒(virus)

Topic#5	firm, market, financial, company, return, investor, investment, bank, stock, model

Topic#22	反應(reaction), 分子(molecule), 高分子(polymer), 結構(structure), 合成(synthesize)
	化合物(compound), 錯合物(complex), 具有(have), 形成(form), 利用(utilize)

Topic#46	市場(market), 報酬(return), 投資(investment), 股票(stock), 指數(index)
	股價(stock price), 交易(transaction), 模型(model), 公司(company), 價格(price)

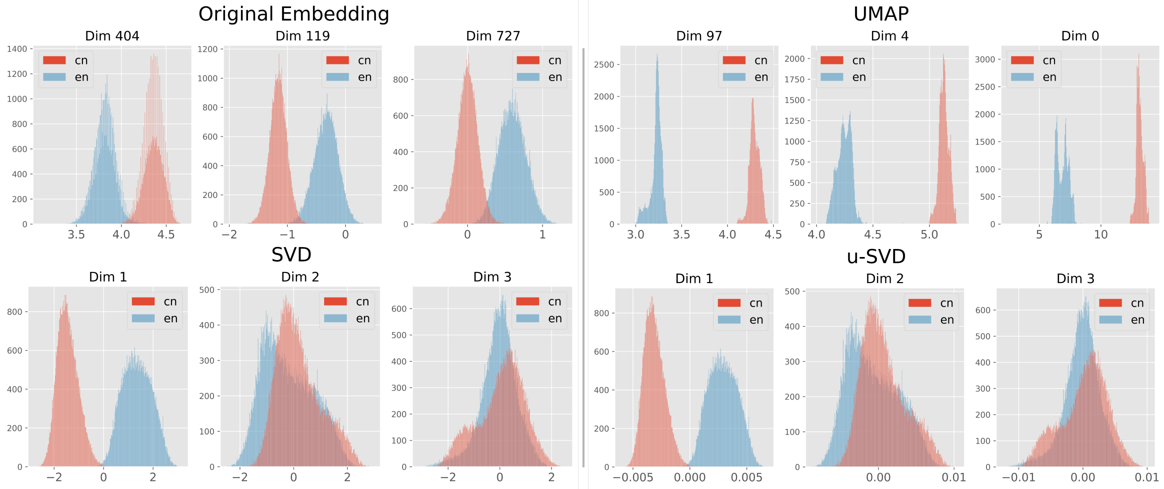


Figure 2: Top 3 language-dependent dimensions, sorted by t-statistic values, for original embeddings and embeddings reduced using UMAP, SVD and u-SVD. We utilize the Cohere multilingual model (see Section 3.2) to encode the documents in one of our experimental datasets, namely ECNews. The value distributions for Chinese (cn) and English (en) documents are indicated by red and blue, respectively. All UMAP, SVD, and u-SVD reduced the dimension size of the original representations from 768 to 100. Appendix A presents the same analysis to the other dataset, namely Rakuten Amazon.

LDDs adversely affect the subsequent document clustering process, as they disproportionately influence the distance calculations between documents during clustering. As a result, LDDs cause the algorithm to cluster documents by language rather than by their semantic meaning. In order to mitigate the negative impacts of LDDs, we repurpose the step 2 of the pipeline from a dimension reduction component to a dimension refinement component. Our dimension refinement component incorporates SVD, leveraging its notable feature that the reduced dimensions are orthogonal to one another. Note that previous researches have long applied SVD for topic modeling (Deerwester et al., 1990; Crain et al., 2012), yet its usage has been confined to monolingual topic modeling for decomposing the term-document matrix to capture the latent semantic structure. We propose a novel approach that applies SVD to neutralize LDDs from the representations generated by MLMs and further reduces the influence of languages. Owing to the orthogonal decomposition property of SVD, when one dimension retains language information, the remaining dimensions are more likely to capture other types of information. The lower-left subplot of Fig. 2 demonstrates that SVD consolidates the scattered LDDs into a concentrated set of reduced dimensions.

We explore two implementations of dimension refinement components, namely unscaled SVD (u-SVD) and SVD with Language dimension Removal (SVD-LR). Both u-SVD and SVD-LR methods follow the same decomposition manner as the standard SVD, which is represented by $E = U\Sigma V^T$. However, unlike standard SVD, u-SVD only utilizes $U \in R^{m\times r}$ to represent m documents in r reduced dimensions. Since U is an orthonormal matrix, u-SVD reduces the influence of LDDs by ensuring that each dimension has a unit length. For instance, the lower-right subplot shows that u-SVD represents the dimensions using smaller scale (see x-axis) compared to the SVD in the lower-left subplot. By reducing the scale of dimensions, u-SVD decreases the negative contributions of LDDs in the subsequent clustering. u-SVD is a conservative approach as it reconciles the effects of LDDs without removing any dimension. In contrast, SVD-LR is more aggressive by removing the most influential LDD after performing SVD. Specifically, we represent the documents using $U\Sigma \in R^{m \times r}$ and use the two-sample t-test to identify the most influential LDD $\hat{r}$, which has the largest difference in the mean values of two languages. Then, SVD-LR removes $\hat{r}$ from $U\Sigma$.

Algorithm 1 Updated Pipeline for Cross-lingual Clustering-based Topic Model

```
Require: MLM, corpus, number of reduced dimensions $r$, number of topics $K$
 1: Obtain $E$ by embedding the corpus using the assigned MLM
 2: $U, \Sigma, V^T = \text{SVD}(E, r)$
 3: if u-SVD then
 4:     $E^* = U$
 5: else if SVD-LR then
 6:     Identify the most influential LDD $\hat{r}$ using two-sample t-test
 7:     Obtain $E^*$ by removing $\hat{r}$ from $U\Sigma$
 8: end if
 9: $C_1, C_2, ..., C_K = \text{Kmeans}(E^*, K)$
10: $\phi_1, \phi_2, ..., \phi_K = \text{c-Tf-IDF}(C_1, C_2, ..., C_K)$
11: return $\phi_1, \phi_2, ..., \phi_K$
```

Algorithm 1 presents the updated pipeline, which is detailed as follows: (1) in line 1, documents are embedded using the MLM to obtain document representations E, (2) from line 2 to line 8, we perform the dimension refinement step [2] using either u-SVD or SVD-LR to obtain refined document representations E^*, (3) in line 9, Kmeans algorithm [3] are applied on E^* to group documents into K topic clusters, and (4) in line 10, we summarize and reconstruct the topic-word distribution for each topic cluster using c-TF-IDF (Eq. 1).

3 Experimental Setup

3.1 Dataset

We conduct experiments using three datasets: (1) **Airiti Thesis** which consists of 163,150 pairs of English and Chinese thesis abstracts (Chang et al., 2020). On average, each abstract contains 165 words. (2) **ECNews** comprises 50,000 Chinese news and 46,850 English news articles, with an average length of 11 words per article. (3) **Rakuten Amazon** is a compilation of 25,000 Japanese and 25,000 English product reviews, with an average of 27 words per review. ECNews and Rakuten Amazon were used in the previous research for cross-lingual topic evaluation (Wu et al., 2023). Considering that ECNews and Rakuten Amazon primarily contain short documents, we include Airiti Thesis in our experiments to evaluate the performance on identifying topics in longer documents.

3.2 Multilingual Language Model

We evaluate our proposed methods and compare them with other methods using three different MLMs: (1) **mBERT** (Devlin et al., 2019) has been investigated for its capability on cross-lingual classification tasks (Pires et al., 2019b). We use transformers[4] to load bert-base-multilingual-cased[5] and use output of special classification token ([CLS]) to get the mBERT embedding for a document. (2) **Distilled XLM-R** (Reimers and Gurevych, 2020) is designed for embedding a paragraph and is based XLM-R (Conneau et al., 2020), which is superior than mBERT in parallel sentence retrieval (Libovický et al., 2020). We use sentence-transformers[6] to access Distilled XLM-R (paraphrase-xlm-r-multilingual-v1). (3) **Cohere multilingual model** has shown its capabilities in various cross-lingual retrieval tasks (Kamalloo et al., 2023). We use the Cohere multilingual model (embed-multilingual-v2.0) by the API[7].

3.3 Baseline & Competitor

We compare three alternative baselines to show the effectiveness of using u-SVD and SVD-LR as dimension refinement step: (1) **original embedding**, referred as *OE*, which is simply generated

[2] We use the SVD implementation from Dask package `https://www.dask.org`.

[3] We use Kmeans implementation with default parameters from scikit-learn package `https://scikit-learn.org/`.

[4] `https://github.com/huggingface/transformers`

[5] `https://huggingface.co/bert-base-multilingual-cased`

[6] `https://www.sbert.net`

[7] `https://txt.cohere.com/multilingual/`

from the given MLM, (2) **UMAP** [8], which is the popular dimension reduction method, whose effectiveness in identifying monolingual topics has been shown (i.e., CETopic) (Zhang et al., 2022), and (3) **pure SVD**, which is used as a benchmark to compare against u-SVD and SVD-LR. Moreover, we compare two recent cross-lingual topic models: (1) **Cb-CLTM** (Chang and Hwang, 2021) incorporates a cross-lingual word space into the generative process of latent Dirichlet allocation (Blei et al., 2003). Cb-CLTM demonstrates its superior performances compared to other probabilistic cross-lingual topic models. To enable the Cb-CLTM, we use pre-aligned English-Chinese and English-Japanese word spaces from MUSE project[9]. (2) **InfoCTM** (Wu et al., 2023) is a neural topic model that identifies topics across languages based on the guidance of the given bilingual dictionary. InfoCTM is the state-of-the-art neural cross-lingual topic model. We follow the report of the InfoCTM to use a Chinese-English dictionary from MDBG[10] and Japanese-English dictionary from MUSE project to link topics across languages.

3.4 Evaluation Metric

We measure the generated topics using two metrics widely adopted in previous CLTMs: CNPMI and Diversity. For each topic $k \in K$, we select top-N represented words for l_1 and l_2 languages, denoted as $\mathcal{W}^{l_1}_{k,N}$ and $\mathcal{W}^{l_2}_{k,N}$.

CNPMI (Hao and Paul, 2020; Chang and Hwang, 2021; Wu et al., 2023) measures the coherence of generated topic words across languages:

$$-\frac{1}{N^2} \sum_{w_i \in \mathcal{W}^{l_1}_{k,N}, w_j \in \mathcal{W}^{l_2}_{k,N}} \frac{\log \frac{Pr(w_i, w_j)}{Pr(w_i)Pr(w_j)}}{\log Pr(w_i, w_j)}, \quad (2)$$

where $Pr(w_i, w_j)$ is the co-occurring probability of words w_i and w_j and $Pr(w_i)$ is the marginal probability of w_i. For Airiti Thesis, we estimate the probability using the comparable abstracts in the Airiti Thesis. For ECNews and Rakuten, we measure the probability using comparable Wikipedia corpus [11]. The CNPMI ranges from -1 (least coherent) to 1 (most coherent), and we report the average CNPMI scores across K topics.

Diversity (Dieng et al., 2020) measures the uniqueness of generated topic words across K topics:

$$\frac{|\bigcup_{1 \le k \le K} \mathcal{W}^{l_1}_{k,N}| + |\bigcup_{1 \le k \le K} \mathcal{W}^{l_2}_{k,N}|}{K \times 2 \times N}, \quad (3)$$

which ranges between 0 (the least diversity) and 1 (the highest diversity). To combine the two aspects, we further compute **Topic Quality (TQ)** (Dieng et al., 2020) as the product of max(0, CNPMI) and Diversity, providing a cohesive measure for our analysis. Note that positive CNPMI contributes to TQ because NPMI measurement positively correlates with human interpretability (Lau et al., 2014). The topic with negative CNPMI are considered to be uninterpretable.

We evaluate top 15 words ($N = 15$) of each topic for CNPMI and Diversity. For more robust comparison, we re-run every method five times using different seeds and report the average performance.

4 Results & Analysis

4.1 Performance of Cross-lingual Topic Model

Table 2 shows the performance of different methods on three datasets. We adopt the following settings. Cohere multilingual model is chosen as the MLM, which embeds every document into 768 dimensional representations. All dimension reduction methods reduce the original embedding from 768 to 100 dimensions. The number of topics (clusters) is set to 50 because InfoCTM (Wu et al., 2023) reports performances on this number for both ECNews and Rakuten Amazon.

The results clearly indicate that incorporating a clustering-based topic model pipeline with three baseline embeddings, including original embedding, UMAP, and SVD, does not perform well in terms of CNPMI and Diversity. We also use feature-wise min-max normalization on UMAP, resulting in UMAP-norm. However, UMAP-norm does not enhance performance. Both Cb-CLTM and InfoCTM exhibit high diversity scores. However, when applied to the Airiti dataset, they generate topics with negative CNPMI scores, suggesting that their generated topics are difficult to be interpreted by human (Lau et al., 2014). The pipelines

[8] We use the implementation from umap-learn package https://github.com/lmcinnes/umap.

[9] https://github.com/facebookresearch/MUSE

[10] https://www.mdbg.net/chinese/dictionary?page=cc-cedict

[11] We use the implementation https://github.com/BobXWu/CNPMI from the authors of InfoCTM (Wu et al., 2023).

Table 2: Comparison of topic quality for baselines, competitors, and our proposed methods.

Dataset	**Airiti**			**ECNews**			**Rakuten Amazon**		
Metric	**CNPMI**	**Diversity**	**TQ**	**CNPMI**	**Diversity**	**TQ**	**CNPMI**	**Diversity**	**TQ**
OE	-0.244	0.570	0.000	0.022	0.554	0.012	0.009	0.290	0.003
UMAP	-0.202	0.572	0.000	0.019	0.598	0.011	0.003	0.265	0.001
UMAP-norm	-0.207	0.585	0.000	0.019	0.613	0.012	0.003	0.264	0.001
SVD	-0.251	0.564	0.000	0.026	0.567	0.015	0.009	0.282	0.003
Cb-CLTM	-0.145	**0.941**	0.000	0.021	0.774	0.016	0.008	0.699	0.006
InfoCTM	-0.087	0.917	0.000	0.044	**0.905**	0.040	0.033	**0.856**	**0.028**
SVD-LR	**0.179**	0.571	**0.103**	**0.087**	0.741	0.065	0.032	0.607	0.019
u-SVD	0.171	0.603	**0.103**	0.086	0.823	**0.071**	**0.037**	0.665	0.025

with u-SVD and SVD-LR result in less diverse topics than Cb-CLTM and InfoCTM but have better CNPMI and TQ on the Airiti and ECNews datasets. Moreover, InfoCTM, SVD-LR, and u-SVD reach comparable CNPMI and TQ on the Rakuten Amazon dataset. These results suggest that u-SVD and SVD-LR can generalize to datasets of different lengths.

4.2 Performance on Different MLMs

To test the generalizability of u-SVD and SVD-LR, we evaluate and compare performances on three MLMs, namely mBERT, Distilled XLM-R, and Cohere Multilingual Model, on the Airiti Thesis. All three MLMs generate document embedding with 768 dimensions. To benchmark with the results shown in Table 2, each document embedding is also reduced or refined to 100 dimensions, and the number of topic clusters is set to 50.

Table 3 reveals that when using mBERT, both SVD-LR and u-SVD achieve only marginal improvement, if any, on topic quality compared to other three baselines. This may be attributed to limited cross-lingual capability of mBERT because it is the first generation MLM. On the other hand, with the document representations generated by more capable MLMs, namely Distilled XLM-R and Cohere Multilingual Model, SVD-LR and u-SVD consistently demonstrates their robust performances and generate topic clusters with better topic quality.

4.3 Sensitivity Analysis on the Size of Reduced Embeddings

To better understand u-SVD and SVD-LR, we conduct sensitivity analysis on the size of embeddings. In this analysis, we use all three datasets and fix the number of cluster topics at 50. We reduce the document representations generated by Cohere Multilingual Model from 768 to 100, 200, and 500 to see their influence on the CNPMI.

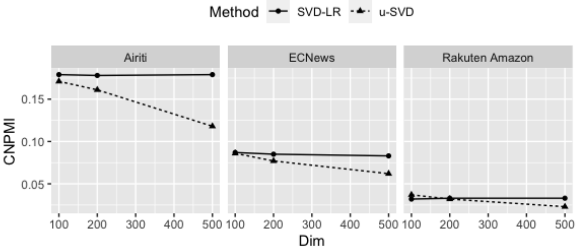


Figure 3: Sensitivity analysis of u-SVD and SVD-LR on different dimensions.

Fig. 3 shows that SVD-LR has a more robust result across different embedding dimensions. SVD-LR preserves the importance weight (i.e., Σ) of each dimension except for the most influential LDD, resulting in robust performance across various dimensions. On the contrary, u-SVD abandons the importance weight of dimensions from SVD to lessen the effect of LDDs. Thus, u-SVD is affected by those dimensions that originally had small singular values, leading to poorer outcomes when more dimensions are utilized. In summary, while both u-SVD and SVD-LR lose some information due to the elimination of LDDs, SVD-LR seems to lose fewer information when more dimensions are introduced.

4.4 Qualitative Result

We apply the Cohere multilingual model to embed the Airiti dataset and use BERTopic (Grootendorst, 2022), which implements the previous pipeline of clustering-based topic model. Table 1 shows the representative words for ten manually sampled topics generated by BERTopic. Each topic consists of top words purely from a single language and is misaligned by the semantic meaning. For in-

Table 3: Topic quality of using three different MLMs.

Method	**mBERT**			**Distilled XLM-R**			**Cohere Multilingual Model**		
	CNPMI	Diversity	TQ	CNPMI	Diversity	TQ	CNPMI	Diversity	TQ
OE	-0.122	0.478	0.000	-0.211	0.600	0.000	-0.244	0.570	0.000
UMAP	-0.190	0.421	0.000	-0.198	0.536	0.000	-0.202	0.572	0.000
SVD	-0.117	0.476	0.000	-0.208	0.580	0.000	-0.251	0.564	0.000
SVD-LR	-0.149	0.492	0.000	0.172	0.527	0.091	**0.179**	0.571	**0.103**
u-SVD	**0.001**	**0.591**	**0.000**	**0.182**	**0.629**	**0.115**	0.171	**0.603**	**0.103**

Table 4: Top representative words of 10 sampled topics from updated pipeline with u-SVD and SVD-LR

u-SVD	
Topic#2	optical, 光學(optics), 雷射(laser), 發光(glow), laser, light, 元件(component), led, 我們(we), 結構(structure)
Topic#7	影像(image), image, 我們(we), 演算法(algorithm), 方法(method), propose, algorithm, 提出(propose), method, video
Topic#9	網路(network), 無線(wireless), 傳輸(transmission), 通訊(communication), network, 我們(we), 使用(use), 系統(system), 提出(propose), propose
Topic#20	polymer, 高分子(polymer), 材料(material), surface, film, increase, high, property, 結構(structure)
Topic#21	投資(investment), 市場(market), 報酬(return), market, return, 股票(stock), 交易(transaction), 指數(index), stock, 投資人(investor)
SVD-LR	
Topic#1	optical, 發光(glow), 光學(optics), 雷射(laser), 元件(component), led, laser, light, 結構(structure), 我們(we)
Topic#6	影像(image), image, 我們(we), 演算法(algorithm), 方法, propose, algorithm, 提出(propose), method, video
Topic#12	polymer, 高分子(polymer), 材料(material), surface, 表面(surface), film, 結構(structure), increase, high, material
Topic#17	網路(network), 無線(wireless), network, 傳輸(transmission), 我們(we), 使用(use), 節點(node), 通訊(communication), 提出(propose), 服務(service)
Topic#21	投資(investment), 市場(market), market, 報酬(return), return, 指數(index), 交易(transaction), 股票(stock), stock, investor

stance, topics #1 & #2 discuss the same topic but are separated into two topics. Table 4 uses the same setting as Table 1 but apply u-SVD and SVD-LR for dimension refinement. Most topics contain representative words across languages and are grouped by the semantic meanings of topics. For example, the concept of "Financial Market" is separated into two topics in Table 1, namely topics #5 & #46, based on languages. On the contrary, as shown in Table 4, topic #21 from u-SVD and topic #21 from SVD-LR include the words of different languages yet with similar concept.

5 Related Work

5.1 Clustering-based Topic Model

Recent works (Sia et al., 2020; Zhang et al., 2022; Grootendorst, 2022) have explored methods that cluster contextualized representations to identify topics from a corpus. Sia et al. (2020) used the BERT model to encode each token into a representation, averaging these representations to obtain a document-level representation. They then applied K-means clustering to these document representations and reconstructed the topic-word distributions using a tf-idf weighting scheme. The coherence performance of their resultant topics was comparable to that of the traditional topic model, LDA (Blei et al., 2003). Similarly, Zhang et al. (2022) and Grootendorst (2022) proposed a pipeline consisting of four steps. First, they used language models, such as sentence BERT (SBERT), to encode documents into representations. Next, they applied the dimension reduction technique UMAP to these representations. In the third step, they used K-means clustering on the reduced representations to generate document clusters, each considered a topic cluster. Finally, they employed a word importance ranking method, c-Tf-IDF, to identify representative topic words. Their pipelines outperformed neural topic models in terms of both efficiency and topic quality. However, the proposed pipeline hasn't been evaluated in cross-lingual settings. Our study aims to fill this gap.

5.2 Language-dependent Component

Several studies (Libovický et al., 2020; Zhao et al., 2021; Chang and Hwang, 2021) have shown that MLM-generated representations contain language-dependent components (LDDs), which signal language identity and hinder cross-lingual transfer. To mitigate such LDDs, Libovický et al. (2020) noted that representations of the same language are closely located in the space. They recommend removing the language-specific mean from the mBERT representations as a solution. However, even after this adjustment, the resulting representations can still be utilized as features to predict the language accurately, suggesting that simply removing the language-specific means from the representations is insufficient. Zhao et al. (2021) propose a method that requires parallel corpus to fine-tune mBERT and XLM-R for generating language-agnostic representations. The method fine-tunes the language model to align the sentence pairs from the parallel corpus. To further close the gap between languages, the method also constrains the representations of different languages to be distributed with zero mean and unit variance. Such an idea is close to our proposed u-SVD; however, u-SVD is a more efficient and appropriate method for models with ample parameters because it does not require parallel corpus and fine-tuning. Chang and Hwang (2021) observed that LDDs prevent their topic model from identifying topics across languages. They proposed training a logistic regression to identify the contributed dimensions (i.e., LDDs) for language identity and removed them from the representations. They found that removing the LDDs helped identify more cross-lingual topics. However, removing the LDDs directly from the original representations comes with the cost of losing semantic completeness. Our SVD-LR eases this issue because utilizing SVD helps us to consolidate the scattered language-dependent dimensions into one specific dimension. Therefore, SVD-LR only removes the most contributed LDD, potentially minimizing the risk of losing other semantic meanings.

6 Conclusion

We investigate the problem with the current pipeline of clustering-based topic model when applied on multilingual corpus, which is caused by language-dependent dimensions in the multilingual contextualized embedding. To solve this problem, we propose two methods for dimension refinement, namely u-SVD and SVD-LR. Our experiments suggest that the updated pipeline with our proposed refinement component is effective in cross-lingual topic identification and results in more coherent topics than existing cross-lingual topic models.

Limitations

This study only evaluates proposed dimension refinement components, u-SVD and SVD-LR, on three MLMs, namely mBERT, XLM-R, and Cohere Multilingual Model. We chose these three MLMs because of their extensive investigations in cross-lingual retrieval tasks. The future work may investigate more other MLMs such as LASER[12], Universal Sentence Encoder[13], and OpenAI embedding API [14]. Extensive experiments on more language pairs are another future work since we only evaluate two English-Chinese datasets and one English-Japanese dataset. It is worth noting that our proposed methods are effective in language pairs from distant and different language families. Furthermore, it's also crucial to investigate our methods for datasets with more than two languages, such as EuroParl.

References

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(Jan):993–1022.

Chia-Hsuan Chang and San-Yih Hwang. 2021. A word embedding-based approach to cross-lingual topic

[12] https://github.com/facebookresearch/LASER

[13] https://www.kaggle.com/models/google/universal-sentence-encoder/

[14] https://platform.openai.com/docs/api-reference/embeddings

modeling. *Knowledge and Information Systems*, 63(6):1529–1555.

Chia-Hsuan Chang, San-Yih Hwang, and Tou-Hsiang Xui. 2018. Incorporating Word Embedding into Cross-Lingual Topic Modeling. In *2018 IEEE International Congress on Big Data (BigData Congress)*, pages 17–24, San Francisco, CA, USA. IEEE.

Chia-Ming Chang, Chia-Hsuan Chang, and San-Yih Hwang. 2020. Employing word mover's distance for cross-lingual plagiarized text detection. In *Proceedings of the Association for Information Science and Technology*, volume 57, page e229.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

Steven P. Crain, Ke Zhou, Shuang-Hong Yang, and Hongyuan Zha. 2012. Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 129–161. Springer US, Boston, MA.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

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, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Adji B Dieng, Francisco J R Ruiz, and David M Blei. 2020. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Philipp Dufter and Hinrich Schütze. 2020. Identifying Elements Essential for BERT's Multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Shudong Hao and Michael J Paul. 2020. An Empirical Study on Crosslingual Transfer in Probabilistic Topic Models. *Comput. Linguist.*, 46(1):95–134.

Yuening Hu, Ke Zhai, Vladimir Eidelman, and Jordan Boyd-Graber. 2014. Polylingual Tree-Based Topic Models for Translation Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1166–1176, Baltimore, Maryland. Association for Computational Linguistics.

Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, Nandan Thakur, David Alfonso-Hermelo, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Evaluating Embedding APIs for Information Retrieval.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 2 - EMNLP '09*, volume 2, page 880, Singapore. Association for Computational Linguistics.

Tiziano Piccardi and Robert West. 2021. Crosslingual Topic Modeling with WikiPDA. In *Proceedings of the Web Conference 2021*, pages 3032–3041, Ljubljana Slovenia. ACM.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019a. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019b. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, Chaoqun Liu, Liang-Ming Pan, and Anh Tuan Luu. 2023. InfoCTM: A mutual information maximization perspective of cross-lingual topic modeling. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, volume 37, pages 13763–13771. AAAI Press.

Michelle Yuan, Benjamin Van Durme, and Jordan L Ying. 2018. Multilingual Anchoring: Interactive Topic Modeling and Alignment Across Languages. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing Language-Agnostic Multilingual Representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

A Two-sample t-test on MLM embeddings for Rakuten Amazon dataset

We use the same setting as in Fig 2 to display the top three language-dependent dimensions of the Rakuten Amazon dataset in Fig 4.

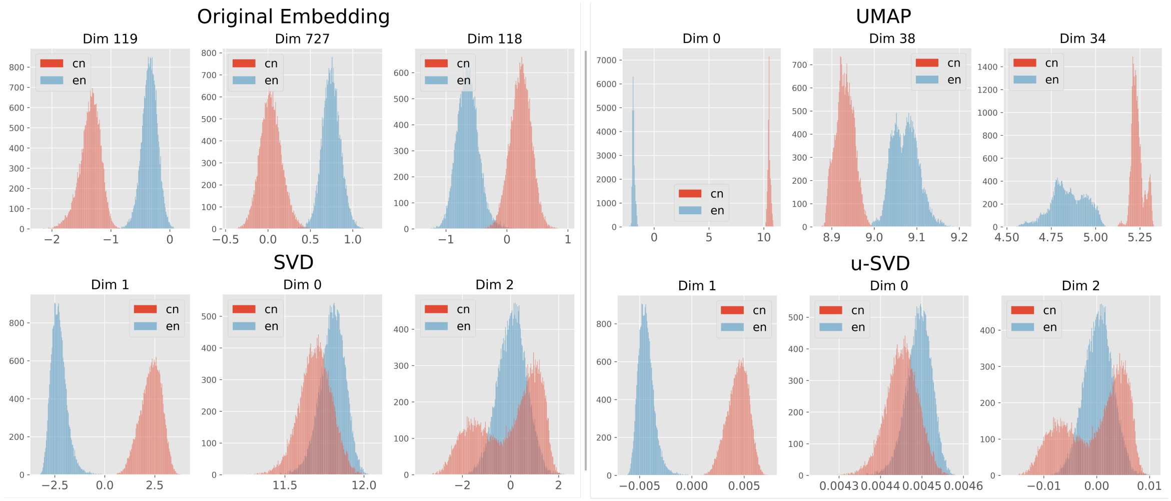


Figure 4: Top 3 language-dependent dimensions, sorted by t-statistic values, for original embeddings and embeddings reduced using UMAP, SVD and u-SVD on Rakuten Amazon dataset.

The Role of Handling Attributive Nouns in Improving Chinese-To-English Machine Translation

Haohao (Lisa) Wang
Carnegie Mellon University
lisaw2@andrew.cmu.edu

Adam Meyers
New York University
meyers@cs.nyu.edu

John E. Ortega
Northeastern University
j.ortega@northeastern.edu

Rodolfo Zevallos
Barcelona Supercomputing Center
rodolfo.zevallos@bsc.es

Abstract

Translating between languages with drastically different grammatical conventions poses challenges, not just for human interpreters but also for machine translation systems. In this work, we specifically target the translation challenges posed by attributive nouns in Chinese, which frequently cause ambiguities in English translation. By manually inserting the omitted particle 的 ('DE'). In news article titles from the Penn Chinese Discourse Treebank, we developed a targeted dataset to fine-tune Hugging Face Chinese to English translation models, specifically improving how this critical function word is handled. This focused approach not only complements the broader strategies suggested by previous studies but also offers a practical enhancement by specifically addressing a common error type in Chinese-English translation.

1 Introduction

The development of Machine Translation (MT) systems for languages with significantly different grammatical structures presents unique challenges (Zhang et al., 2024), particularly in the treatment of function words, which may be implicitly understood in one language but require explicit translation in another. In Chinese, for example, the absence of attributive particles such as 的 ('DE') can lead to ambiguities and inaccuracies in English translations. Nowadays, MT systems use large training datasets, general-domain datasets or extracting parallel texts from existing corpora for domain-specific tuning (Devlin et al., 2018; Liu et al., 2019; Conneau et al., 2020; Chi, 2021; Kocmi et al., 2022). However, these approaches often fall short when addressing nuanced linguistic features that are domain-specific or underrepresented in available training datasets.

学生问题　(without DE)
学生的问题　(with DE)

Figure 1: Two Translations of *Student Question*

In this work, we adopt a focused approach to improve the translation of Chinese attributive nouns to English, targeting the challenges posed by the implicit usage of the particle 'DE'. Our research begins with a feasibility test of function words in Chinese, including prepositions, conjunctions, particles, and modals. We particularly examine the impact on translation quality when these words are omitted. We identified 'DE,' which explicitly signifies adjectives in Chinese–examples in Figure 1. With "DE", the phrase refers to an NP (a question a student asks or a question assigned to a student).Without "DE", the phrase can also be a sentence with the approximate meaning "a student asks a question". "DE" turns out to be a critical function word whose absence significantly affects the clarity and accuracy of translations. We hypothesize that tuning machine translation models specifically on use cases of attributive nouns can improve their performance. To confirm this, we create a parallel set of data using news article titles from the Penn Chinese Discourse Treebank (Xue, 2005). This dataset includes the original Chinese titles and their modified English translations, where 'DE' was manually inserted to accurately reflect its implied usage. We use this dataset to fine-tune Chinese-English MNT models. Then we conduct a rigorous evaluation using a sample of 1,000 sentences from the UM Chinese English parallel corpus afterwards. The results showed notable improvements in BLEU (Papineni et al., 2002) and CHRF (Popović, 2015) scores.

Our contributions are twofold: (i) We developed a novel, focused parallel corpus specifically addressing the translation challenges posed by attributive nouns in Chinese, thereby enhancing the semantic accuracy of translated texts. (ii) We fine-tune NMT models using our innovative dataset to refine the translation of complex grammatical struc-

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 57–61
January 20, 2025. ©2025 Association for Computational Linguistics

tures from Chinese to English. This approach not only improves translation accuracy but also contributes to the broader understanding of function word impact in machine translation, especially for languages with substantial grammatical differences.

2 Related Work

Our research proves largely relevant to the foundational ideas presented by Heylen et al. (1994), who explored the concept of lexical functions as cross-linguistic semantic primitives essential for crafting translation strategies. These strategies aim to preserve the semantic integrity of collocations across languages, highlighting the profound impact that precise modeling of function words can have on translation accuracy. This seminal work sets the stage for understanding how nuanced handling of function words is critical to maintaining meaning across different languages. Building on these insights, subsequent research has further underscored the significance of function words in translation. Notably, Zhang et al. (2017) developed advanced embedding techniques to integrate the nuanced usages of function words directly into the translation process. Their approaches—concatenation, partitioning, and usage-specific embeddings—are designed to enhance the NMT system's understanding of function words, which is vital for achieving accurate translations. Additionally, Kuo (2019) identified discrepancies in the usage of function words between machine-translated and original Chinese texts, noting that the overuse of frequent function words could lead to "translationese." This phenomenon occurs when the translated text retains too many features of the source language, underscoring the need for nuanced handling of function words in translation models. Further research by He et al. (2019) delved into how function words affect the performance of neural machine translation (NMT) systems. Their findings demonstrate the importance of adequately addressing function words within NMT systems to improve translation outcomes. These collective research efforts illustrate the evolving understanding and importance of function words in machine translation, emphasizing that effective handling of these elements is essential for bridging the gap between languages with diverse grammatical structures. However, while these studies focus on the issue of properly handling function words when they are present, they often overlook the effect these words can have when they are absent but implied. This is particularly relevant in the case of attributive nouns in Chinese, which frequently cause ambiguities in English translation. Complementing these studies, our research specifically targets the translation challenges posed by these attributive nouns, proposing a method to address the challenge effectively. By developing a targeted dataset and refining the NMT model to recognize and correctly translate implied function words, our work aims to reduce ambiguities and enhance the clarity and accuracy of translations between these linguistically diverse languages.

3 Methodology

This work is driven by the observation that closed-class words, including prepositions, conjunctions, particles, and modals, primarily serve grammatical rather than lexical roles in a language. These elements provide the scaffolding that holds sentences together, determining the syntax and influencing the semantics. Crucially, they are essential for conveying the correct relationships and structures within sentences, directly impacting the logic and intended meaning of the original statements. Capturing the meanings of closed-class words is thus fundamental to preserving the logical structure and intended meaning of sentences in translation.

3.1 Feasibility Test

We identified our functional words of interest based on the following criteria:

1. The word is optional in the original language; that is, deleting the word does not necessarily change the meaning of the sentence.

2. Deleting the word results in a grammatically correct but ambiguous sentence.

3. The translation of the sentence carries a different meaning from the original.

We compiled a list of 82 such words in Chinese. 436 ChatGPT-generated and manually-checked sentences incorporating these function words were generated to create a focused test corpus. To isolate the effect of each function word, we developed a script that systematically removed each word from the sentences, generating a new output file for each word's omission. This method allowed for a controlled examination of the impact on translation quality when these words were absent. After inspecting the output of these files with that of the original file without any of the closed-class words removed through Argos translate, surprisingly, we observed that NMT performed generally well when handling omitted conjunctions given that deleting the conjunctions did not change the meaning of the sentence in the original language, The omitted prepositions had limited impact as well, which can be attributed to a key characteristic of the Chinese language, where the presence of certain auxiliary words can compensate for the omission of prepositions, thereby reducing the impact on translation accuracy and coherence. Omitting particles results in ungrammatical sentences in the original language, making them unsuitable objects according to our test criteria. Lastly, among the modals, the only case where all of the above criteria are met is "DE", which is a particle that typically follows an adjective and is omitted in the case of an attributive noun.

3.2 Data Development

Realizing that use cases of attributive nouns in Chinese cause ambiguity and inaccuracy when translated into English, we extracted the titles of all the articles from the Penn Chinese Discourse Treebank, noting that titles of news articles are a common places for the use of attributive nouns. We ran Argos Translate on the list of titles to obtain their English translation, then manually inserted the implied particle "DE" back into the titles where the use of attributive nouns occurred and ran Argos Translate on the modified list again.

After comparing the translation results before and after modification, out of 165 titles, 143 contained uses of attributive nouns, and 135 of these showed improvements in translation quality in terms of coherence, completeness, and accuracy. This manual process resulted in a parallel dataset with the original Chinese titles and their English translations after modification, ready to be used as tuning data for the MT models we prepared to verify if targeted interventions on function word handling can enhance MT system performance.

4 Experiment Results

To test our test-dataset, we decided to fine-tune some of the most prominent English-Chinese NMT models. The NMT models were fine-tuned with 2 subsets: 60 and 1k sentences from the approximately 67.5k sentences categorized under the News section of UM Corpus. This was done to assess the effectiveness of our approach and determine if more exhaustive testing of our dataset was necessary.

4.1 Models

We use the following models for fine-tuning.

- MarianNMT (Junczys-Dowmunt et al., 2018): is an open-source neural machine translation framework built upon the Transformer architecture. It supports various architectures such as Transformer, Transformer-Base, and Transformer-Big. The Transformer architecture consists of self-attention mechanisms and feed-forward neural networks, allowing for parallel computation and capturing long-range dependencies in the input sequence. MarianNMT is known for its efficiency, scalability, and ease of training on custom datasets. It allows for quick experimentation with different configurations and is widely used in both research and production settings.

- NLLB-200 (Costa-jussà et al., 2022): stands for Neural Language Lattice Based model, a novel approach specifically designed for low-resource languages. Unlike traditional NMT models, NLLB-200 utilizes a lattice-based decoding mechanism, which helps to handle the ambiguity often present in low-resource language translation tasks. This model incorporates techniques to efficiently capture linguistic nuances and improve translation quality even with limited training data.

- mBART (Liu et al., 2020): is a multilingual variant of BART (Bidirectional and Auto-Regressive Transformers), a transformer-based model specifically designed for sequence-to-sequence tasks like machine translation. mBART leverages a pretraining phase where it learns to generate text in multiple languages simultaneously. This multilingual pretraining enables mBART to effectively transfer knowledge across languages, making it particularly useful for multilingual translation tasks. Additionally, mBART introduces a shared tokenization scheme across languages, facilitating direct comparison and transfer of information between different language pairs.

5 Results

The fine-tuning process demonstrated a slight but noticeable improvement in BLEU scores and precision metrics, particularly when the sample size was increased to 1000 sentences. This larger sample size helped to better showcase the efficacy of the targeted interventions. Out of the evaluated sentences, 45 showed improved translation quality while 42 exhibited some regression. However, the magnitude and quality of improvements were substantially more significant than the regressions, underscoring the potential benefits of fine-tuning MT systems with carefully curated data.

6 Discussion

6.1 Interpretation of Results

Our experiments provide valuable insights into the role of closed-class words in machine translation, particularly from Chinese to English. Initial testing and fine-tuning of the Helsinki NLP model with a dataset specifically enhanced for attributive nouns showed nuanced but measurable improvements in translation accuracy. The slight increase in BLEU scores and precision metrics, especially noticeable after expanding the sample size, suggests that targeted interventions on specific linguistic features, such as attributive nouns, can significantly enhance machine translation system performance. This improvement is critical because attributive nouns in Chinese often omit particles like DE, leading to significant ambiguities when translated into English without proper contextual handling.

Our findings emphasize the importance of accurately translating closed-class words. These words, while not carrying separate meanings themselves, crucially structure the syntax and semantics of sentences. Incorrect translation or omission can disrupt the logical flow of the translated text, resulting in outputs that are grammatically correct but semantically flawed. Moreover, the improved handling of these words through manual adjustments and model tuning illustrates that even minor enhancements in linguistic structure treatment can substantially improve the clarity and coherence of translated texts. This is particularly vital for languages like Chinese, where the omission of specific function words is a common practice and poses significant challenges in automated translation contexts.

6.2 Limitations and Future Research Directions

While the improvements are encouraging, the relatively modest increases in BLEU scores highlight the limitations of current machine translation technologies in handling complex linguistic phenomena. The necessity for manual intervention in creating the tuning dataset also reveals a significant gap in automated systems' ability to comprehend and reproduce nuanced linguistic features independently. Initially, we attempted to use a script (using Python's Jieba package) to automatically identify nouns by tags and append the missing particle "DE." However, due to the multifunctional nature of many Chinese words, which can serve as nouns, verbs, or adjectives, this script did not perform adequately, leading us to manually insert "DE" in the original text to ensure data accuracy. This manual process was time-consuming and significantly limited the corpus's scope, thus constraining our ability to produce and observe the impact of a larger dataset. Future research should explore the development of more sophisticated NLP models that incorporate deeper linguistic analyses into the translation process, potentially reducing the need for manual adjustments. Expanding the scope of the study to include other languages with similar linguistic features could further our understanding of the universal applicability of our approach. Additionally, investigating the use of advanced machine learning techniques such as deep learning and neural networks may provide new avenues to automate the identification and correct translation of closed-class words more effectively, potentially leading to significant advancements in the field, especially for languages with complex syntactical structures.

7 Conclusion

In conclusion, our study demonstrates the potential benefits of targeted linguistic interventions in machine translation. By focusing on the accurate translation of closed-class words, especially in languages like Chinese where these words play a crucial grammatical role, we can significantly enhance the semantic accuracy of translated texts. Continued exploration and refinement of these techniques are essential for advancing the capabilities of machine translation technologies in the coming years.

Acknowledgments

We extend our gratitude to the institutions that provided invaluable support. Our affiliates include New York University, Northeastern University, Pace University, Carnegie Mellon University, Universitat Pompeu Fabra, and the Barcelona Computing Center.

Model	Pre-trained		Fine-Tuned - 60 sent		Fine-Tuned - 1k sent	
	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF
MarianNMT	37.8	70.6	37.8	70.6	38.2	72.1
NLLB-200	38.3	72.1	38.3	72.2	39.5	75.1
mBART	36.7	64.8	36.7	64.8	37.1	66.4

Table 1: BLEU and CHRF results of the 3 NMTs used to measure the performance of our test-dataset

References

Z Chi. 2021. mt6: Multilingual pretrained text-to-text transformer with translation pairs. *arXiv preprint arXiv:2104.08692*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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*.

Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael R Lyu, and Shuming Shi. 2019. Towards understanding neural machine translation with word importance. *arXiv preprint arXiv:1909.00326*.

Dirk Heylen, Kerry G Maxwell, and Marc Verhagen. 1994. Lexical functions and machine translation. *arXiv preprint cmp-lg/9410009*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chen-li Kuo. 2019. Function words in statistical machine-translated chinese and original chinese: A study into the translationese of machine translation systems. *Digital Scholarship in the Humanities*, 34(4):752–771.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Nianwen Xue. 2005. Annotating discourse connectives in the Chinese treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 84–91, Ann Arbor, Michigan. Association for Computational Linguistics.

Jinyi Zhang, Ke Su, Haowei Li, Jiannan Mao, Ye Tian, Feng Wen, Chong Guo, and Tadahiro Matsumoto. 2024. Neural machine translation for low-resource languages from a chinese-centric perspective: A survey. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Can a Neural Model Guide Fieldwork? A Case Study on Morphological Data Collection

Aso Mahmudi[϶] **Borja Herce**[Ʒ] **Demian Inostroza Améstica**[϶]
Andreas Scherbakov[϶] **Eduard Hovy**[϶] **Ekaterina Vylomova**[϶]
[϶]The University of Melbourne [Ʒ]University of Zurich
amahmudi@student.unimelb.edu.au vylomovae@unimelb.edu.au

Abstract

Linguistic fieldwork is an important component in language documentation and the creation of comprehensive linguistic corpora. Despite its significance, the process is often lengthy, exhaustive, and time-consuming. This paper presents a novel model that guides a linguist during the fieldwork and accounts for the dynamics of linguist-speaker interactions. We introduce a novel framework that evaluates the efficiency of various sampling strategies for obtaining morphological data and assesses the effectiveness of state-of-the-art neural models in generalising morphological structures. Our experiments highlight two key strategies for improving the efficiency: (1) increasing the diversity of annotated data by uniform sampling among the cells of the paradigm tables, and (2) using model confidence as a guide to enhance positive interaction by providing reliable predictions during annotation.

1 Introduction

According to UNESCO, around 2,000 languages are currently classified as endangered and over half of the languages spoken today might disappear by the end of the century.[1] In 2022, the organisation has declared the start of the decade of indigenous languages, and many linguists increased their efforts in documentation and revitalisation. But language documentation is a drawn-out, iterative, and exhausting process. A linguist would normally visit a language community several times to interview speakers and collect the data. During each visit, she or he would focus on tasks such as elicitation of words and language rules by offering them questionnaires or asking them to tell stories. Between visits, the linguist would focus on processing, revising the data, and forming working linguistic hypotheses that will be further revised during the next face-to-face sessions. The amount of time spent in interaction with speakers is an important limiting resource, as native speakers often get tired in lengthy sessions, leading to a decline in their attention and interest, and, as a result, in poorer data quality (Bowern, 2015).

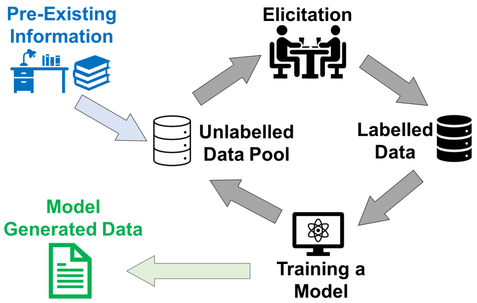


Figure 1: Illustration of the proposed word elicitation process model.

In this paper, we introduce **a neural system that guides the linguist, making the process of data collection more efficient**.[2] The proposed model takes into account pre-collected data, identifies potential gaps in it, and informs the linguist of the (most informative) parts that should be collected in the next iteration. In contrast to existing approaches, we for the first time incorporate a measure that reflects an important ergonomic aspect of linguist-speaker interactions: we explicitly distinguish the following two cases of "atomic" linguist-to-speaker interactions: (1) either a linguist makes a correct guess satisfying the speaker, or (2) seeks more information (e.g., upon producing ungrammatical utterances). The latter action tires the informant more than the former. Therefore, assuming that much greater cost associated to case (2) compared to case (1), we frame the planning of interaction sequences as an optimisation task.

As a case study, we focus on morphological in-

[1] https://www.un.org/development/desa/indigenouspeoples/indigenous-languages.html

[2] You can find all the code for this paper at https://github.com/Aso-UniMelb/neural-fieldwork-guide

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 62–72
January 20, 2025. ©2025 Association for Computational Linguistics

flection data as it is characterised by high regularity and systematicity (Vylomova, 2018) and neural models are particularly good at capturing regular patterns in data and have previously demonstrated high accuracy on morphological inflection shared tasks (Cotterell et al., 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021; Kodner et al., 2022; Goldman et al., 2023). As we aim to identify more data-efficient approaches, we also provide a comparative analysis of a variety of sampling strategies (1) under a variety of data conditions as well as (2) in terms of their relevance and utility for the fieldwork pipeline. For the first aspect, we include typologically diverse languages representing major morphological processes (fusion, agglutination), a variety of morphological complexities, and with ranging amounts of data available. For the second, we evaluate the models' ability to capture paradigm cell inter-predictability (discussed in Section 4.2).

Our main contributions are:

1. A novel approach to evaluate neural models that takes into account the nature of linguist-speaker interactions;

2. Evaluation of state-of-the-art models and sampling approaches for data-efficiency and ability to capture inter-cell predictability.

2 Background

2.1 Motivation for the Word-and-Paradigm Model

A key task in linguistic data collection involves the development and management of interlinear glossed texts, where morphological forms are broken down into units that carry meaning. While tools like "FieldWorks Language Explorer (FLEx)"[3] offer some semi-automated assistance, interlinear glossing remains a highly time-intensive task for field linguists. The SIGMORPHON 2023 shared task on interlinear glossing (Ginn et al., 2023) highlighted efforts to automate this process and demonstrated that the availability of morphological segmentation plays a crucial role in achieving high accuracy. Still, morphological segmentation itself is a non-trivial task and a complicated problem in computational morphology (Batsuren et al., 2022a).

An alternative method for morphological annotation is to adopt a model which does not necessitate segmentation. Copot et al. (2022) also recommend a word-based approach to morphological annotation, especially for under-resourced and under-described languages. When working on a new language, a linguist collects and analyses wordforms, making generalisations about their relationships, and trying to identify morphological organisation, i.e., the structure and the size of the morphological paradigm (the number of paradigm cells). Having the paradigm structure, the linguist can then study the inter-predictability of the paradigm cells, trying to identify **principal parts**, the minimal subset of paradigm cells that provides all the necessary information to generate the other cells within the paradigm (Finkel and Stump, 2007). In the well-known case of Latin, for example, all forms of the verb can be generated from just 4 forms (Finkel and Stump, 2009). Such knowledge allows for a more compact representation of linguistic rules and higher efficiency in data collection.

Many typical tasks in morphology such as paradigm discovery (Erdmann et al., 2020a), paradigm completion (Durrett and DeNero, 2013), paradigm cell filling problem (Ackerman et al., 2009), and morphological inflection (Kodner et al., 2022) are often approached using a word-based model. In theoretical linguistics, the Word-and-Paradigm model (Blevins, 2016) offers a foundational framework for this word-based approach.

2.2 Making the Data Collection Process More Efficient

What is the best strategy to collect language data? As this process is time-consuming, it is essential to increase its efficiency. We explore active learning approaches in this paper. **Active Learning (AL)** has a well-established history in different NLP tasks (Zhang et al., 2022) and fits well with the language documentation process, where field linguists periodically consult with informants. For instance, Palmer (2009) used AL for real fieldwork experiments of a morpheme labelling task with two native speakers by examining three sequential, random, and uncertainty sampling strategies. Muradoglu and Hulden (2022) studied the simulated AL for a morphological inflection task on different languages with different sampling strategies. Muradoglu et al. (2024) found that the success of an inflection model on a test set largely depends on the entropy of the edit operations (required to transform a lemma into a target form) in the training data, and higher entropy which can be obtained by a uniform sampling across paradigm cells tends

[3] https://software.sil.org/fieldworks/

to improve the model's performance. Erdmann et al. (2020b) proposed an approach to automate the paradigm cell filling problem task by manually providing a few forms. However, their method is impractical in real fieldwork settings because it requires the speaker (oracle) to frequently review the entire paradigm table.

3 A Model of the Word Elicitation Process

Word Elicitation is a technique used in linguistics to gather lexical and morphosyntactic data from native speakers with minimal contextual information. While corpora show what people *say*, elicitation uncovers what *can be said* (Meakins et al., 2018). To discover the morphological features, linguists usually change one feature at a time (Bowern, 2015). Elicitation cannot be sustained for an extended period in fieldwork, so it is recommended to limit it to around 20 hours spread across multiple sessions (Abbi, 2001). In each session, the speaker is asked carefully designed short questions, and the linguist analyses the responses to generalise potential patterns.

This study focuses on modelling word elicitation during morphological data collection (as is illustrated in Figure 1), with an emphasis on optimising process efficiency.

3.1 Main Task and Initial Assumptions

The task involves filling in all plausible cells of the paradigm tables with correct inflected word forms. Cells that do not apply to specific lemmas are excluded from the process.

We assume the availability of pre-existing data, either gathered during early fieldwork stages or sourced from previous descriptive resources.

This data should include:

1) a basic word list (similar to the Swadesh list) consisting of verbs, nouns, adjectives, and other parts of speech provided in their dictionary forms (lemmas), and

2) a range of morphosyntactic features for each part of speech, which may be derived from prior studies or inferred from closely related languages, where applicable. We assume the knowledge of possible morphosyntactic feature combinations (tagsets such as "`N;ACC;PL`").

3.2 Linguist–Speaker Interactions

We now turn to the model of linguist-speaker interactions during the word elicitation process in morphological data collection. We model a native speaker as an oracle system that has access to complete paradigms for all lemmas (labelled data pool). As an input, it receives (1) a lemma and (2) a target feature combination (tags corresponding to a paradigm cell).[4] The linguist model is a neural system that can send requests to the speaker model. The requests might come at a certain cost as the process of word elicitation is exhausting, especially for native speakers (Bowern, 2015). Whenever the linguist model retrieves a form or makes an incorrect prediction (in both cases the speaker model needs to return a valid form), it gets a penalty score of 1. In the case the linguist model checks a form and it is correct, the speaker is satisfied, and the linguist model does not get any penalty score. Hence, the linguist model has to optimise the retrieval process in order to minimise the penalty and increase the prediction accuracy.

At some point, the linguist has to decide to stop the data collection process and return to their office. This means that they assume that the collected data is informative enough to accurately predict all the missing parts. Hence, at the final step, the linguist model predicts all the missing cells for each lemma. Whenever the prediction is incorrect, the model receives a penalty of 1 as well.

3.3 The Data Collection Model

Once the initial data described in Section 3.1 is prepared, the linguist model generates for each lemma in the word list an unlabelled data pool. The pool consists of possible empty cells in the paradigm that correspond to plausible morphosyntactic feature combinations.

As mentioned above, given the potentially large number of forms, it is impractical to ask the speaker model for all of them. Instead, a small subset of cells is selected over several rounds (cycles) of elicitation, and the linguist model is trained to generalise from that subset. The key here is to identify and target the most informative cells early on to gain a better understanding of the morphological structure.

Inspired by the 20-hour elicitation timeframe advised in fieldwork (Abbi, 2001), and assuming 100 items are asked per hour, we limit our interaction to approximately 2,000 speaker (oracle) queries spread over five sessions, with 400 data wordforms retrieved in each cycle.

[4] In this work, we assume some linguistic expertise and knowledge of the features.

Language	Code	Family	Typology	POS	Forms	Lemmas	APS
English	eng	Germanic	analytic	V	5,120	1280	4
Latin	lat	Romance	fusional	V	240,078	5,185	89
Russian	rus	Slavic	fusional	N	208,198	18,008	16
Central Kurdish	ckb	Iranic	fusional	V	21,375	375	57
Turkish	tur	Turkic	agglutinative	V	80,264	380	295
Mongolian	khk	Mongolic	agglutinative	N	14,396	2057	8
Central Pame	pbs	Oto-Manguean	fusional	V	12,528	216	58
Murrinh-patha	mwf	Southern Daly	polysynthetic	V	1,110	30	37

Table 1: Total number of wordforms, lemmas and average paradigm size (APS) for the selected part-of-speech (POS) across examined languages.

In the first cycle, the linguist model has no prior knowledge about the informativeness of each cell for facilitating generalisation and predicting other cells. At this stage, the model may either sample cells uniformly from the pool or start by gathering a few complete paradigms. Note that in the latter option, the number of tables that can be collected from 400 queries will depend on their size in the corresponding language. In some languages such as English, it might cover 100 paradigm tables, while in others, like Turkish, it might represent only two full paradigms (their average verbal paradigm size is greater than 200). Importantly, the availability of complete paradigms allows a linguist to infer cell inter-predictability and estimate the predictive power of each cell in paradigm tables and identify the principal parts. In our experiments, we explore both strategies.

Once the initial processing is complete, the linguist needs to decide on the next cells to request from the speaker. Several strategies can be employed here: only checking the cells the linguist is most confident about (this reduces penalty but might be uninformative), exploring the most informative parts of the paradigm, or retrieving the cells with the highest uncertainty. We employ active learning (Ren et al., 2021) to optimise the sampling process. Each cycle here involves training a neural inflection model (a linguist model) to make generalisations about the data. While neural models typically require large amounts of data for training, they can generate predictions with varying levels of confidence at each training stage. We leverage this evolving capability to streamline interactions.

After several cycles of data collection, when we reach the approximate limit of 2,000 oracle queries, the trained neural model is used to predict the remaining pool data and its accuracy on these final predictions is evaluated.

4 Experimental Setup

4.1 Datasets

For this study, we selected 8 typologically diverse languages: English, Latin, Central Kurdish, Russian, Turkish, Khalkha Mongolian, Central Pame, and Murrinh-patha. The languages range in their morphological organisation, paradigm sizes, and levels of documentation. Table 1 provides a summary of the dataset specifications organised by language. The datasets are derived from UniMorph (Batsuren et al., 2022b) and VeLePa (Herce, 2024, Central Pame). The data samples are presented in the form of triplets consisting of a lemma (e.g., "dog"), a target form ("dogs"), and morphosyntactic tags ("N;PL").

4.2 Experiments

In our simulated data collection procedure, the oracle (speaker) is provided with access to the entire morphological dataset (labelled data pool). Additionally, for the remainder of the process, it also stores the forms that the linguist retrieved along with their predictions (if applicable). The linguist model has access to the data pool excluding the target form (i.e. unlabelled data). The linguist model, using its sampling strategy, selects a subset of lemma-target tag set combinations (a paradigm cell) from the pool and requests the corresponding target forms. When making a request to the oracle, the linguist model includes a predicted form if it has sufficient confidence in the prediction. If the prediction is correct, the oracle does not apply a penalty.

To evaluate sampling strategies and the interaction model, we design four experimental setups, which are described as follows. In all experiments, the labelled data were collected over five cycles of AL, with 400 target forms gathered per cycle. The only exception is Murrinh-patha, where limited data availability required reducing the collection to

100 forms per cycle. Please note that whenever a neural model was trained, it was initialised from scratch and trained using all the data collected up to that point.

Exp. 1: In the first experiment, we model a baseline scenario when a linguist only asks a speaker to provide forms, without any particular strategy to select the most informative ones. Thus, here we uniformly sample a fixed number of cells from the pool in each of the five cycles. No suggestions were provided to the oracle throughout the experiment.

Exp. 2: In the second experiment, the linguist still does not have any particular sampling strategy but after the initial session, the linguist can make predictions with varying degrees of confidence based on observations from previous sessions and suggests the confident predictions to the speaker (hence reducing the chances of penalty). We modelled this case by using uniform sampling for each cycle and training a neural model on the collected data to provide confident predictions. The model predicted forms for all cells in the pool to determine an average confidence level. Subsequently, it retrieved the forms of randomly selected samples from the oracle and passed a prediction if its confidence surpasses the average confidence level.

Exp. 3: In the third experiment, a linguist collects some data, then studies it, and tries to fill in all the remaining cells in the whole data pool. Then they check with the speaker the forms they are most confident about and ask the speaker to provide forms they are puzzled about. This experiment follows a similar approach to the second, where a model was trained after the first cycle using random sampling. However, in the subsequent cycles, the sampling strategy was not random. The model generated predictions for the remaining pool data and ranked them based on confidence. Predictions with the highest confidence were queried from the oracle accompanied by a prediction, while the least confident predictions were obtained without one.

Exp. 4: The fourth experiment illustrates a scenario where the linguist first asks the speaker to complete full paradigms for a few lemmas. Then, the linguist assesses the inter-predictability of the cells to focus primarily on the cells with higher predictive power. We describe this experiment in more detail as it introduces a novel method not previously explored. In the first cycle, the linguist model selects a small list of lemmas and asks the oracle for their complete paradigm table. The number of lemmas depends on the average size of the paradigm per language (assuming approximately 400 forms were queried). These data are used to identify the inter-predictability of cells in the paradigm tables.

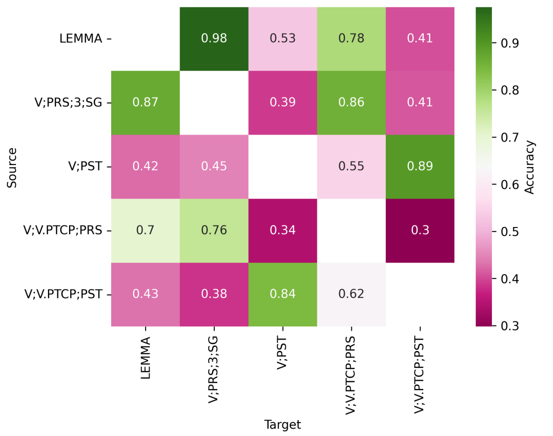


Figure 2: A heatmap showing the accuracy of predictions for English verbs.

We illustrate this process using English verbal paradigms due to its relatively small size. If we exclude the syncretic and non-morphologically realised forms, English paradigm tables would contain one lemma (the infinitive) and four inflected forms (present tense third person singular, simple past, past and present participle). Thus, we retrieve 400 English forms by requesting 100 paradigm tables, generate a dataset of all 2,000 possible reinflection permutations (20 for each of the 100 verbs) and divide it into training, development, and test sets, with 45%, 45%, and 10% of the data in each set, respectively. To explore the inter-predictability of cells, only once before the second cycle, we train a neural re-inflection model (details in Appendix A) considering each cell as a source, aiming to predict from it the remaining forms in the corresponding paradigm table. We consider all possible source–target cell combinations, e.g. "`went + V;PST + V;PRS;3;SG`" was used as the input and "`goes`" as the output of the model to measure the predictability of "`V;PST`" with respect to "`V;PRS;3;SG`" for the lemma "`go`". Figure 2 shows a heatmap that indicates the model accuracy on the test set for different source and target tag combinations. The heatmap reveals that, in English, the lemma is generally a more informative source for predicting third-person singular present tense ("`V;PRS;3;SG`") and present participle ("`V;PTCP;PRS`") forms, compared to past tense ("`V;PST`") or past participle ("`V;PTCP;PST`") forms. Additionally, there is greater inter-predictability be-

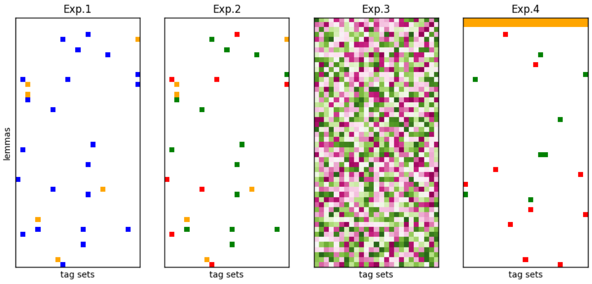


Figure 3: A simplified overview of sampling strategies used in the second cycle of the experiments. Blue cells represent samples retrieved without any predictions or confidence checks. Dark green cells denote confident ones retrieved with predictions, while dark red cells indicate low confidence cells with no predictions sent to the oracle. Orange cells indicate those that were selected in the first cycle and removed from the pool.

tween simple past tense and past participle forms. The predictive power of an individual cell can be estimated from the average accuracy across the target cells. The system did not rely only on the most predictive cell. Instead, it employed these weights as fuzzy values in a weighted random sampling process. Based on these estimations, the system assigned weights for the remaining cells of the pool.

The sampling strategy for the following cycles of Exp.4 was similar to Exp.2, with the key difference being that in the second experiment, the sampling was uniform whereas in the fourth it was weighted random. The weights were determined by the estimated predictive power of each tagset. Like in Exp.2, a model was trained to predict the wordforms, and its predictions were passed to the oracle if the model had higher confidence in them.

To summarise the differences between the experiments, consider the second cycle illustrated in Figure 3. In Exp.1, cells were randomly selected for retrieval without any prediction. In Exp.2, the model passed predictions for confident cells, while no predictions for low confidence cells. In Exp.3, the most confident predictions were selected for retrieval with prediction, while the least confident ones were retrieved without prediction. Exp.4 followed a similar approach to Exp.2 but gave higher selection priority to more informative cells.

5 Evaluation

We evaluate the performance across the four experiments in terms of the accuracy of the final model and the efficiency of the process, using the following measures:

Accuracy on unseen data After the final cycle of the AL process, we calculate the accuracy of the inflection model trained on all retrieved samples in predicting the target form for the remaining samples in the pool (considering it as the test set).

Normalised Efficiency Score We define a penalty score as an integer number by summing the number of times we call the oracle (excluding the times we propose a correct guess for the target form) and the number of incorrect predictions of the final model on the unseen test set. Since the size of the datasets is not the same, we normalised the penalty by the total number of forms per language. To better capture the efficiency of the elicitation process, we introduce a new metric—the complement of the normalised penalty—referred to as the Normalised Efficiency Score (NES). This score is calculated as follows:

$$NES = 1 - \frac{P_1 + P_2 + P_3}{N} \tag{1}$$

where P_1 is the number of forms retrieved from the oracle without a suggestion, P_2 is the number of forms retrieved with an incorrect suggestion, P_3 is the number of incorrect predictions in the final test set, and N is the total number of target forms in the dataset.

6 Results and Discussion

We conducted evaluation of the four experiments described in Section 4.2, across all the languages in our datasets. For each iteration of active learning, the data labelled by the oracle was split into 90% for training and 10% for development. This data was used to train an inflection model from scratch using a neural character-level transformer, following the hyper-parameters from Wu et al. (2021). At the end of each experiment, all remaining data in the pool was used as the test set and the final model predicted the corresponding target forms.

6.1 Model Accuracy

Table 2 provides the target form prediction accuracy on the test set (the remaining samples in the pool) of examined languages. Among the various sampling strategies tested in our experiments—uniform sampling, weighted random sampling based on estimated inter-predictability values, and sampling based on the model's confidence—uniform sampling yielded the highest prediction accuracy. Our findings are consistent with

lang	Exp.1	Exp.2	Exp.3	Exp.4
tur	**98.2**	97.6	93.5	95.7
ckb	97.5	**97.6**	90.3	95.5
eng	89.2	89.0	89.0	**90.9**
khk	83.3	**85.1**	77.8	84.9
rus	84.2	**85.8**	71.1	84.3
lat	**72.3**	71.3	49.1	67.3
pbs	72.2	**73.8**	62.9	64.7
mwf	**80.0**	78.4	62.1	79.6
Average	84.6	**84.8**	74.5	82.9

Table 2: Accuracy of the final model on remaining pool after the final cycle. Experiments 1 and 2 used identical sampling and their results are almost equal according to this evaluation metric.

lang	Exp.1	Exp.2	Exp.3	Exp.4
tur	95.8	**96.3**	92.5	94.1
ckb	88.4	**92.4**	87.0	90.3
eng	54.2	68.7	**72.9**	66.1
khk	71.7	**78.2**	72.0	76.4
rus	83.4	**85.2**	70.9	83.7
lat	**71.7**	70.9	49.2	66.9
pbs	60.7	**66.0**	58.2	57.3
mwf	44.0	**54.4**	48.3	49.6
Average	71.2	**76.5**	68.9	73.2

Table 3: Normalised Efficiency Score of each experiment on different languages.

previous studies (Muradoglu and Hulden, 2022; Muradoglu, 2024), confirming that random sampling across all paradigm cells is an effective strategy that cannot be outperformed easily when using smaller amounts of data, demonstrating its efficiency in the elicitation process.

Next, we analyse the model's performance across active learning cycles. In all experiments, approximately 2,000 forms (500 for Murrinh-patha) were retrieved in total. Figure 4 shows the accuracy of the inflection models on the remaining pool data in each cycle of the experiments. It demonstrates that accuracy improves with each cycle, initially increasing rapidly and then rising more slowly in the later cycles. However, Exp.3 shows limited accuracy gains for languages like Latin, Kurdish, and Russian. These languages have slots in their paradigms that either copy the lemma or exhibit regular consistent inflections. Confidence-based sampling tends to select these slots for providing suggestions, which restricts the diversity of the training data. This limitation is particularly evident in our Latin data, given its larger number of unique lemmas.

Due to the extremely low accuracy in the first cycle of Exp.4, we excluded them from Figure 4. This poor performance can be attributed to the limited lexical diversity of the training data, as most of it comes from just a few paradigm tables. However, in the third cycle, the accuracy in Exp.4, which used a weighted random sampling, improves significantly and approaches the performance of the uniform random sampling used in Exp.1 and Exp.2.

6.2 Interaction Efficiency

We now turn to an analysis of interaction efficiency. We observe that incorporating the confidence values of the inflection model for its predictions leads to sending more accurate predictions to the oracle, further enhancing the process's overall efficiency. Table 3 shows the normalised efficiency score for the experiments per language.

To better understand the interaction efficiency, we analyse the outcomes as follows: The linguist models (except in Exp.1), to minimise penalties, submitted their predictions with queries when sufficiently confident. Nonetheless, these predictions were not always accurate. Figure 5 illustrates the number of data samples retrieved from the oracle, segmented by the correctness of the submitted prediction. Exp.3 outperformed the others by employing a non-random sampling strategy based on the model's confidence. Overall, this demonstrates that, to some extent, we can rely on the model's confidence to enhance the efficiency of the interaction process.

To evaluate the impact of prioritising the completion of a few paradigm tables over the rest of the elicitation process, we designed Exp.4, where cell informativeness within paradigms was estimated and influenced the proportion of data retrieval. However, the results indicate that this approach does not significantly enhance the model's performance or efficiency, as successful generalisation in neural models largely depends on the lexical diversity and entropy of the training data.

7 Conclusion

In this paper, we evaluated neural models in their ability to guide fieldwork by accounting for the nature of linguist–speaker interactions in the process of language documentation. Focusing on morphological data collection, we investigated various strategies for data sampling. Our results showed that uniform random sampling across paradigm cells results in more representative data and yields better generalisation in low-resource scenarios. Furthermore, we discovered that incorporating the model's confidence levels enhances interaction by

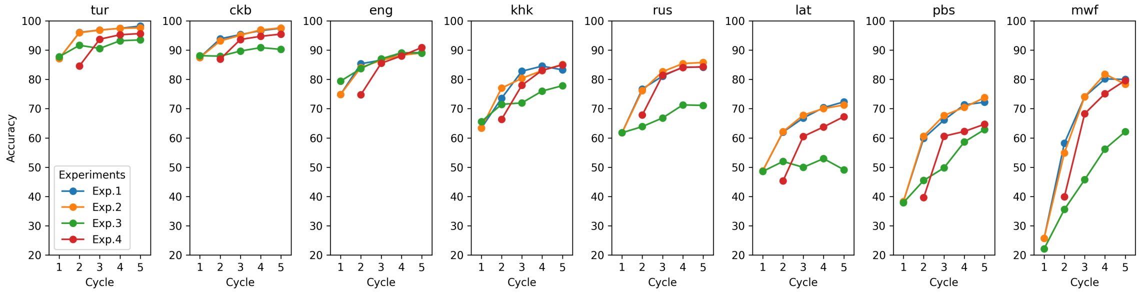


Figure 4: Accuracy on remaining pool data in each cycle of the active learning process for each language.

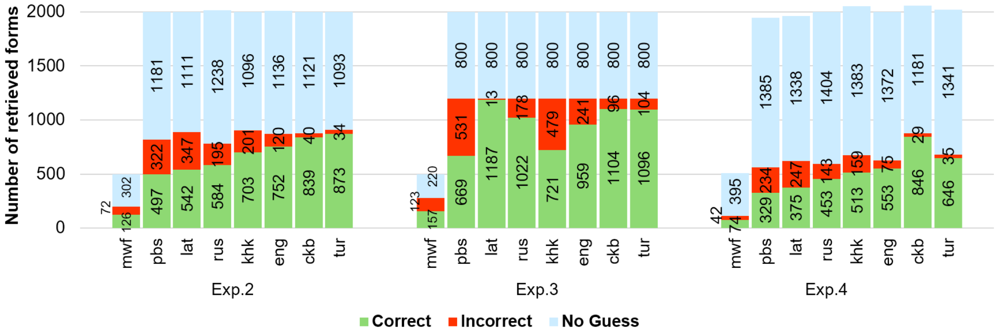


Figure 5: Submitted predictions along the requests to the oracle in each experiment. Exp.1 is omitted as all its requests were without a prediction.

guiding decisions on whether to send a prediction. This approach improves the process by offering predictions as suggestions during data annotation tasks.

8 Future Work

This study employed a simulated active learning approach for morphological data collection. To translate this into a real-world application, two user interfaces would be necessary: one for linguists to input existing data and another one for native speakers to provide the desired information.

Since native speakers may find complex tasks that require linguist knowledge tedious, we suggest that the linguist prepares a variety of simple sentences to change the user interface into fill-in-the-blank tasks. Naturally, designing these sentences is a challenging task that varies for each part of speech and requires some preliminary understanding of the language, which can be informed by the morphosyntactic features collected earlier. During the system's elicitation process, the speaker can fill in or correct the relevant part of the paradigm by considering the context and the lemma. For instance, to elicit the past tense of the verb 'sleep' in English, the prompt could be "`I [sleep] yesterday.`" This approach resembles the SIGMORPHON 2018 shared task 2 (Cotterell et al., 2018).

In addition, to speed up the speaker data entry in the first cycle, the linguist can write some general rules as regular expressions to generate suggestions for each cell. Instead of typing from scratch, the speaker can accept the suggestion or make minor corrections where necessary.

If a required cell is not available for a word, the speaker should let the linguist know through the interface. The cell should be removed from the data pool and should be reviewed by the linguist later. For instance, if a noun is incorrectly labelled as a verb and the system requests its past form, its part of speech should be corrected.

Future studies could explore using inflection classes in evaluation or sampling strategies, though significant challenges remain. Defining the exact number of classes in each language requires considerable granularity, such as determining how many of them would be necessary to accurately predict irregular English verb forms —– a matter on which linguists and educators may disagree. Additionally, resource limitations, especially in low-resource languages lacking comprehensive dictionaries or grammatical descriptions, hinder the identification of inflection classes for all lemmas.

Limitations

We evaluated our method in a simulated manner across a variety of languages with different amounts of available data. We are assuming that our existing data (a wordlist, parts of speech, and morphological tags) are accurate and do not require any modifications during data collection. Additionally, we are assuming that the speaker does not make any errors during data entry. In real-life fieldwork scenarios, any type of error can occur, and a linguist should address them by making corrections as early as possible.

Ethics Statement

We do not foresee any potential risks and harmful use of our work. Our analyses are based on licensed data which are freely available for academic use.

Acknowledgements

This work was supported by 2024ECRG140. We also express our highest gratitude to John Mansfield, Khuyagbaatar Batsuren, and reviewers for their constructive and valuable feedback.

References

Anvita Abbi. 2001. *A Manual of Linguistic Field Work and Structures of Indian Languages*. Lincom GmbH, München.

Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar: Form and Acquisition*, page 0. Oxford University Press.

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. 2022a. 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, 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. 2022b. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

James P Blevins. 2016. *Word and paradigm morphology*. Oxford University Press.

Claire Bowern. 2015. *Linguistic fieldwork: A practical guide*. Springer.

Maria Copot, Sara Court, Noah Diewald, Stephanie Antetomaso, and Micha Elsner. 2022. A Word-and-Paradigm Workflow for Fieldwork Annotation. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 159–169, Dublin, Ireland. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. In *Pro-*

ceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection, pages 1–30, Vancouver. Association for Computational Linguistics.

Greg Durrett and John DeNero. 2013. Supervised Learning of Complete Morphological Paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta, Georgia. Association for Computational Linguistics.

Alexander Erdmann, Micha Elsner, Shijie Wu, Ryan Cotterell, and Nizar Habash. 2020a. The Paradigm Discovery Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7778–7790, Online. Association for Computational Linguistics.

Alexander Erdmann, Tom Kenter, Markus Becker, and Christian Schallhart. 2020b. Frugal paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8248–8273.

Raphael Finkel and Gregory Stump. 2007. Principal parts and morphological typology. *Morphology*, 17(1):39–75.

Raphael Finkel and Gregory Stump. 2009. What your teacher told you is true: Latin verbs have four principal parts. *Digital Humanities Quarterly*, 3(1).

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 Shared Task on Interlinear Glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. SIGMORPHON–UniMorph 2023 Shared Task 0: Typologically Diverse Morphological Inflection. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–125, Toronto, Canada. Association for Computational Linguistics.

Borja Herce. 2024. VeLePa: Central Pame verbal inflection in a quantitative perspective. *Morphology*, 34(3):281–319.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 2022. SIGMORPHON–UniMorph 2022 Shared Task 0: Generalization and Typologically Diverse Morphological Inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Felicity Meakins, Jennifer Green, and Myfany Turpin. 2018. *Understanding linguistic fieldwork*. Routledge.

Saliha Muradoglu. 2024. *Leveraging computational methods for morphological description: A case study of Nen*. PhD Thesis, The Australian National University, Canberra, Australia.

Saliha Muradoglu, Michael Ginn, Miikka Silfverberg, and Mans Hulden. 2024. Resisting the Lure of the Skyline: Grounding Practices in Active Learning for Morphological Inflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 47–55, Bangkok, Thailand. Association for Computational Linguistics.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. How to choose data for morphological inflection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7294–7303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexis Mary Palmer. 2009. *Semi-automated annotation and active learning for language documentation*. Ph.D. thesis, The University of Texas at Austin.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 Shared Task on Morphological Reinflection: Generalization Across Languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A Survey of Deep Active Learning. *ACM Computing Surveys*, 54(9):180:1–180:40.

Ekaterina Vylomova. 2018. *Compositional Morphology Through Deep Learning*. PhD Thesis, The University of Melbourne.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the Transformer to Character-level Transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A Survey of Active Learning for Natural Language Processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Model details

You can find all the code associated with this paper at `https://github.com/Aso-UniMelb/neural-fieldwork-guide`. The implementation and setup details of the neural architectures used in this study are provided below for clarity and reproducibility.

1) Re-inflection Models (used only in Exp.4): These models are one-layer Bidirectional Long Short-Term Memory (BiLSTM) networks implemented using PyTorch. The key hyperparameters used for training are:

- Batch size: 16
- Hidden dimension: 256
- Learning rate: 0.005
- Training duration: 20 epochs

The training process utilises a specific method for embedding morphosyntactic tags. Instead of embedding each tag individually, the tags for each data sample are embedded as a single unit. This method ensures compact representations. The source tag set, input word, and target tag set are then encoded into a dense vector representation.

2) Inflection Models (All Experiments): A neural character-level transformer architecture was employed to train the inflection models used across all experiments. This architecture follows the hyperparameters detailed in Wu et al. (2021). Transformers are particularly suited for this task due to their ability to capture long-range dependencies and complex relationships in inflection data. The character-level approach ensures a fine-grained understanding of morphological patterns at the subword level.

Comparable Corpora: Opportunities for New Research Directions

Kenneth Church
Northeastern University
k.church@northeastern.edu

Abstract

Most conference papers present new results, but this paper will focus more on opportunities for the audience to make their own contributions. This paper is intended to challenge the community to think more broadly about what we can do with comparable corpora. We will start with a review of the history, and then suggest new directions for future research.

1 Introduction

The success of chat bots in many languages demonstrates the power of comparable corpora (CC) and pivoting via English. We will start with a review of the history of CC, and then suggest new directions for future research:

1. More depth: CC are normally used for simple tasks such as bilingual lexicon induction (BLI), but CC can be used for much more interesting views of lexical semantics.
2. Compare and Contrast: CC are normally used to make simple comparisons over language pairs, but they can be used for contrasts as well as comparisons (in monolingual settings as well as multilingual settings).
3. More modalities: Now that vectors encode everything (text in many languages, pictures, audio, video), we can compare and contrast everything with everything.
4. Bursting filter bubbles: bots made in America are trained on corpora from an American perspective with American biases. We should not impose American values on others.

2 Historical Background

2.1 Parallel Corpora

Table 1 shows some examples of parallel corpora. HF and LDC in Table 1 refer to HuggingFace[1] and the Linguistic Data Consortium,[2] respectively. More parallel corpora can be found on HF by searching for *parallel*, *aligned* and *translation*.

The main application of parallel corpora has been machine translation (Brown et al., 1993). Shannon's noisy channel model (Shannon, 1948) was originally motivated for applications in communication (telephones), but it has been used for many other applications including machine translation. That is, to translate from English to French, one imagines that French speakers think in English, like English speakers do, but for some reason, when French speakers talk, the noisy channel converts their English to French.

$$E \rightarrow Noisy\ Channel \rightarrow F \tag{1}$$

The task of the translation system is to recover the original English, E, from the observed French, F. These days, it has become standard practice to use

Resource	Cites
Europarl (Koehn, 2005)	4634
OPUS (Tiedemann, 2012)	2255
HF: Helsinki-NLP/opus-100	
(Resnik and Smith, 2003)	848
MultiUN (Eisele and Chen, 2010)	327
HF: Helsinki-NLP/multiun	
Bible (Pratap et al., 2024)	254
(Akerman et al., 2023)	
HF: Flux9665/BibleMMS	
LDC: Hansard French/English	
LDC: Hong Kong Hansards	
HF: NilanE/ParallelFiction-Ja_En-100k	
HF: sentence-transformers/parallel-sentences	
HF: tiagoblima/bible-ptbr-gun-gub-aligned	
HF: dsfsi/vukuzenzele-sentence-aligned	
(Marivate et al., 2023)	

Table 1: Examples of parallel corpora

[1] https://huggingface.co/

[2] https://www.ldc.upenn.edu/

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 73–82
January 20, 2025. ©2025 Association for Computational Linguistics

English	French	Sense
bank	banque	money
	banc	river
duty	droit	tax
	devoir	obligation
drug	médicament	medical
	drogue	illicit
land	terre	property
	pays	country
language	langue	medium
	langage	style
position	position	place
	poste	job
sentence	peine	judicial
	phrase	grammatical

Table 2: Using Hansards for Word Sense Disambiguation (WSD), based on Table 2 in Gale et al. (1992)

neural networks for translation, but it used to be popular to use Hidden Markov Models (HMMs) to find the most likely English, $\hat{E}$, based on a prior (language model), $Pr(E)$, and a bilingual dictionary, $Pr(F|E)$.

$$\hat{E} = \text{argmax}_E Pr(E) Pr(F|E) \qquad (2)$$

Much of the discussion below will focus on the bilingual lexicon, $Pr(F|E)$. $Pr(E)$, the language model in Eqn (2), is relatively well estimated because we can re-use monolingual LLMs that have been developed for other applications. The bilingual lexicon, $Pr(F|E)$, assigns probabilities to all sequences of English, E, and French, F. It was standard practice, at least at first, to estimate $Pr(F|E)$ from parallel corpora such as the English-French Canadian Hansards.

The rest of this section on history will largely focus on the lexicon. After introducing comparable corpora as an alternative to parallel corpora, we will motivate WSD (word-sense disambiguation). Much of the research on WSD started with bilingual word-senses, but it should be noted that word-senses are different in monolingual and bilingual dictionaries.

There has also been considerable work on transferring monolingual lexical resources such as WordNet and VAD to more languages. Unfortunately, much of this work uses translation to pivot out of English in inappropriate ways, as we will see.

This section will end with a review of BLI (bilingual lexicon induction). The BLI literature uses more modern methods in machine learning than previous methods for inducing lexicons from CC, but BLI benchmarks (such as MUSE) may not be as effective as older WSD methods for addressing classic challenges with translations of ambiguous words. A classic example is *bank*, which is translated as *banque* and *banc* in the Canadian Hansards, depending on the sense. Unfortunately, this ambiguity is not captured in the MUSE benchmark where *bank* translates to *banque* (but not *banc*). In the reverse direction, MUSE has translations for both *banque* and *banc*, but they translate to different English words, *bank* and *bench*, respectively. Comparisons of ambiguities in Hansards (Table 2) and MUSE (Table 5 and Table 6) suggest that MUSE is not testing WSD as much as the older literature. Another concern with MUSE is that most words in the benchmark translate to themselves. These concerns suggest that there may be room to introduce a new benchmark that would make a stronger case for comparable corpora (CC).

After discussing history, the next section will discuss more radical challenges for the future: lexical semantics, transfer learning, filter bubbles and connections between academic search and CC.

2.2 Comparable Corpora (CC)

The term, *comparable corpora*, was introduced in (Fung and Church, 1994; Rapp, 1995; Fung and Yee, 1998; Fung, 2000) to address limitations with parallel corpora. Parallel corpora are available for a few genres such as parliamentary debates (Hansards) and religion (Bible), as shown in Table 1. Since most texts and most genres are not translated, we can collect larger and more diverse corpora if we relax the restriction on translation. CC replace a single parallel corpus with two monolingual corpora, ideally on similar (comparable) topics.

2.3 Word-Sense Disambiguation (WSD)

In addition to machine translation applications mentioned above, parallel corpora have also been used to disambiguate ambiguous words such as *bank*, as illustrated in Table 2. Bar-Hillel (1960) thought machine translation was impossible when he could not figure out how to disambiguate words such as those in Table 2. It was obvious that the translation depends on a solution to WSD.

Gale et al. (1992) used this argument in reverse to obtain large quantities of labeled text for WSD research. They used parallel corpora such as Hansards to find instances of ambiguous words such as *bank*, and use the French translations to label each instance of *bank* as either "money" sense or "river" sense. After labeling the English in this way, they threw away the French and used the sense-labeled text to train and test machine learning methods for WSD.

2.4 Monolingual Senses != Bilingual Senses

This approach was successful in reviving interest in WSD research, though it should be mentioned that bilingual lexicography is different from monolingual lexicography. Consider the word *interest*. This word has many senses including a "money" sense and a "love" sense, among others. A monolingual dictionary will describe each of these senses in considerable detail. However, there will be little to say about *interest* in an English-French bilingual dictionary because the same complications are shared between the English word and its French equivalent. Thus, the approach above is more effective for words like those in Table 2 where the word is ambiguous in one language but not the other, and less effective for words like *interest*, which are equally ambiguous in both languages.

2.5 Inappropriate Uses of Translation

Parallel corpora are limited in a number of ways. Genre is perhaps the most obvious limitation, but a more serious limitation may be distortions introduced by translation.

When I was first working with Hansards in the 1990s, I tried to pitch parallel corpora to Sue Atkins, a lexicographer who specialized in English-French bilingual dictionaries. She rejected my pitch, objecting to "translationese"[3] as "unnatural" natural language. In addition, she criticized concordance tools for parallel corpora because they failed to distinguish source and target languages. Examples of these tools can be found in the sketch engine;[4] these tools show examples of a word in one language as well as its equivalents in other languages.

2.5.1 XNLI: A Multilingual version of NLI

Much of the work on parallel corpora treats the source and target languages as equivalent (with equal status), ignoring distortions introduced by translation. We should be more careful about translation artifacts in many benchmarks. Artetxe et al. (2020) call out XNLI, a English version of an NLI task. The monolingual NLI task depends on word overlaps between the premise and the hypothesis, but many of these crucial overlaps are lost in translation in the XNLI version where premises and hypotheses are translated independently. Too much of the work in computational linguistics uses translation to pivot via English in inappropriate ways.

Synset	French Glosses
dog.n.01	canis_familiaris, chien
cat.n.01	chat
house.n.01	maison
bank.n.01	banque, rive

Table 3: Global WordNet pivots from English

English	Hausa	V	A	D
aaaaaaah	aaaaaaa	0.48	0.61	0.29
aaaah	aaaah	0.52	0.64	0.28
aardvark	ardvark	0.43	0.49	0.44
aback	abin mamaki	0.39	0.41	0.29
abacus	abacus	0.51	0.28	0.49
abalone	abalone	0.50	0.48	0.41

Table 4: NRC-VAD pivots from English using Google Translate; V = Valance, A = Arousal & D = Dominance

2.5.2 No Language Left Behind (NLLB)

Abdulmumin et al. (2024) report serious problems with FLORES (Goyal et al., 2022) in four African languages. A common problem was the use of Google Translate, which sometimes produced "incoherent or unclear" Hausa text. FLORES is an important test set for NLLB (no language left behind) (NLLB Team et al., 2022).

2.5.3 WordNet and VAD

Table 3 and Table 4 show attempts to use translation to pivot from English to other languages. WordNet[5] (Miller, 1995) and NRC-VAD (Mohammad, 2018)[6] were originally designed for English. Translation was used to transfer them to more languages. Note that translation introduces losses; bank.n.01 cannot be both the money sense (*banque*) and the river sense (*rive*). I asked a colleague, a native speaker of Hausa, to comment on Table 4. None of the Hausa words in the table are that useful. Most of the words in the Hausa column are English, with the exception of *abin mamaki* which Google translates

[3] https://en.wiktionary.org/wiki/translatese

[4] https://www.sketchengine.eu/guide/parallel-concordance-searching-translations/

[5] https://www.nltk.org/howto/wordnet.html

[6] https://saifmohammad.com/WebPages/nrc-vad.html

to *what a surprise* in English. My informant did not know what *aback* means in English even though his English is excellent. When I explained it to him, we agreed that this translation is not convincing.

English	French
bank	banque, banques, *but not* banc
duty	devoir, *but not* droit
drug	drogue, médicament
land	terre, terrain, terres, *but not* pays
language	langue, langues, langage
position	position, *but not* post
sentence	peine, phrase, sentence
good	bien, bon, bonne, bonnes, bon
bad	mal, mauvais, mauvaise, bad

Table 5: Some examples from MUSE: fr → en

In short, there are many problems with using translation to pivot from English to many other languages. It is unlikely that the structure of the English WordNet ontology and the English VAD lexicon is universal over all languages. In the West, we slay dragons, but in the East, dragons are good luck. In the West, white is common for weddings and black is common for funerals, but in some places, white is common for funerals, and in other places, red is common for weddings. Even the list of concepts is likely to vary from one language to another. Many of the English words in Table 4 are not (much of) "a thing" in Hausa.

French	English
banc	bench, *but not* bank
banque	bank, banking
droit	right, law
devoir	duty
drogue	drug, drugs, drogue
médicament	medicine, drug, medication
terre	land, earth, soil, terre
terrain	land, terrain
terres	land, lands
langue	language
langues	language, languages
langage	language
position	position
peine	sentence, pain, penalty, sorrow
phrase	sentence, phrase
sentence	sentence, sentencing

Table 6: Some examples from MUSE: en → fr

Dict	Pairs	Src	Tgt	Src=Tgt
en → fr	113,286	94,681	97,035	73,471
fr → en	113,324	97,021	94,730	73,471

Table 7: MUSE Dictionary Sizes

2.6 Bilingual Lexicon Induction (BLI)

Much of the work on BLI is based on the MUSE benchmark [7] (Lample et al., 2017; Conneau et al., 2017). The MUSE benchmark provides:

1. fastText[8] embeddings for 30 languages, and
2. gold set of bilingual dictionaries, $D_{l_i \to l_j}$ for 110 pairs of languages: l_i, l_j. The gold sets are split into training (seed) dictionaries and test dictionaries.

See section 2.2 of (Sharoff et al., 2023) for an introduction to vector space models and CC. The fastText embeddings, $X_l \in \mathbb{R}^{|V_l| \times d}$, contain a row for each word in the vocabulary, V_l, for language l. The rows are vectors of length d, where d is the number of hidden dimensions.

Each dictionary, $D_{l_i \to l_j}$ consist of a list of pairs of words in the two languages. Table 7 counts the number of pairs in both directions, as well as the number of unique words in the source language (src) and target language (tgt). Many of the pairs use the same word in both languages, as indicated by the last column.

The task is to estimate a dictionary, $\hat{D}_{l_i \to l_j}$, for a pair of languages, l_i and l_j. We then compare estimates, $\hat{D}$, with gold dictionaries, D. A simple approach is to use the training (seed) dictionaries to estimate a rotation matrix, $R \in \mathbb{R}^{d \times d}$, where $R = \text{argmin}_R \, ||RX_{l_i} - X_{l_j}||_F^2$. It is standard practice to estimate R with the orthogonal Procrustes problem[9] (Schönemann, 1966).

At inference time, we start with a vector in X_{l_i}, and then rotate those vectors by R and use approximate nearest neighbors (ANN) (Bruch, 2024) to find nearby vectors in X_{l_j}.

Early work on CCs attempted to collect word lists (Kilgarriff et al., 2014) and infer bilingual lexicons; MUSE updates this approach using modern methods in machine learning. That said, MUSE may not be as effective as older methods for WSD because of gaps. Examples from $D_{\text{fr}\to\text{en}}$ and

[7] https://github.com/facebookresearch/MUSE

[8] https://github.com/facebookresearch/fastText

[9] https://docs.scipy.org/doc/scipy/reference/generated/scipy.linalg.orthogonal_procrustes.html

$D_{\text{en}\to\text{fr}}$ are shown in Tables 5-6; some of the ambiguities in Table 2 are covered, and some are not. An example of a gap is: *bank* (en) → *banc* (fr); this pair is missing from both $D_{\text{en}\to\text{fr}}$ and $D_{\text{fr}\to\text{en}}$.

3 Challenges for the Future

3.1 BLI, PMI and Lexical Semantics

Much of the work on BLI uses a simple view of a bilingual lexicon where single words in one language correspond to single words in another language, more or less one-for-one. Obviously, the relationship is far more complicated than this. The phrasal verb, *ask for*, is similar to *request*, violating the one word for one word assumption.

3.1.1 Etymology

More seriously, there is a difference in register, going back to the Norman Conquest in 1066. For a few hundred years after 1066, the English Court spoke French. As a result, English borrowed many words from French. The French term typically has a higher register than the older English equivalent; the peasants raise *cows, calf* and *swine* so the aristocracy can eat *beef, veal* and *pork.*[10]

3.1.2 Distributional Methods

Much of the work on BLI does not take advantage of etymology because work on BLI is based on the Distributional Hypothesis[11] (Harris, 1964) and Firth's "You shall know a word by the company it keeps" (Firth, 1957). The distributional hypothesis is convenient for computation, suggesting "(unlabeled) corpora are all we need," though many aspects of linguistics go beyond distributional evidence, e.g., etymology, lexical semantics.

There are interesting connections between popular distributional methods, e.g., PMI (pointwise mutual information), Word2Vec and LLMs (large language models). The connection between BLI and Word2Vec was mentioned above. Levy and Goldberg (2014) view Word2Vec as a factored representation of PMI (Church and Hanks, 1990). BERT and chat bots can be viewed as an enhancement of Word2Vec; instead of representing words as vectors, we now represent sequences of 512-subword units as vectors.

Relation	PMI	Lexical Sem	Back Trans
Synonyms	large	$=$ (equiv. rel.)	large
Antonyms	large	$\neq$ (anti-sym)	small
Is-a	small	$\leq$ (partial order)	small
Part-Whole	large		small

Table 8: PMI $\neq$ Lexical Semantics

3.1.3 Lexical Semantics

As mentioned above, lexical semantics is a challenge for distributional methods. While there are some similarities between PMI (collocations) and lexical semantics (synonyms, antonyms, is-a), there are also some important differences, as shown in Table 8. PMI scores are large when words appear near one another more than chance. Consequently, both synonyms and antonyms have large PMI scores because documents often compare and contrast this with that. Similarly, PMI scores can be large for other words that appear near one another, e.g., *window, door* and *house*. Large PMI scores do not necessarily imply synonymy.

Back translations are also mentioned in Table 8. Back translations are more effective than PMI for distinguishing synonyms from antonyms. If we take a random walk over MUSE dictionaries and start from *good*, such walks will often take us to synonyms, but rarely to antonyms. There is an opportunity to propose a theory of translation and collocation based on linear algebra and graph theory. This theory should explain the observations in Table 8 where antonyms are close in terms of PMI but not in terms of random walks on translations.

3.1.4 Avoid Pivoting via English

As mentioned in subsection 2.4, monolingual lexicography is different from bilingual lexicography. For example, *interest* has many senses in monolingual dictionaries, but not in bilingual dictionaries. *Bank* is ambiguous in English, but not in French. Bilingual dictionaries become interesting when the senses are not isomorphic. Table 3 and Table 4 take an overly simplistic view of the structure of the lexicon where the ontology (and VAD values) are assumed to be universal. Translating from English is likely to introduce distortions. Can we do better than pivoting via English?

3.2 Transfer Learning

Suppose we want to transfer from a high resource language such as English to growth opportunities such as Indonesian (id) and Hausa (ha). We prefer

[10] https://www.csmonitor.com/The-Culture/In-a-Word/2021/0510/They-re-cows-in-the-field-but-beef-on-the-table

[11] https://aclweb.org/aclwiki/Distributional_Hypothesis

Language	Wikipedia	Joshi	S2 Abstracts	ACL	HF Datasets	HF Models	Speakers
en	6,917,939	5	88,348,938	103,000	10,749	50,717	1456M
zh	1,452,669	5	3,061,847	71,800	1202	4495	1138M
hi	163,524	4	2,848	8,740	421	1388	610M
es	1,992,685	5	2,742,468	28,600	945	3245	559M
fr	2,650,236	5	2,772,266	35,500	1064	4033	310M
id	711,624	3	2,234,953	4,230	395	1317	290M
ar	1,625,651	5	149,043	17,900	558	1681	274M
bn	160,408	3	445	3,270	298	788	273M
pt	1,138,923	4	1,937.959	9,660	596	1935	264M
ru	2,012,648	4	509,503	13,300	799	2307	255M
ur	215,081	3	454	3,220	204	658	232M
de	2,964,125	5	1,227,473	42,900	789	348	133M
ja	1,438,806	1	317,394	38,200	596	2887	123M
mr	98,559	2	275	1,480	193	642	99M
te	101,681	1	13	2,120	223	589	96M
tr	624,742	4	370,727	8,490	398	1389	90M
ta	169,766	3	728	3,980	263	1030	87M
vi	1,294,281	4	44,477	3,010	474	1188	86M
tl	47,891	3	933	1,100	116	451	83M
ko	691,121	4	793,921	16,900	534	2741	82M
ha	51,659	2		823	98	441	79M
jv	74,159	1		535	76	342	68M
it	1,893,522	4	184,535	14,400	516	2129	68M
gu	30,474	1	23	263	174	581	62M
th	169,192	3	41,628	12,700	326	900	61M
kn	33,026	1	143	1540	178	534	59M
am	15,374	2	96	1110	117	493	58M
yo	34,080	2	18	799	123	458	46M

Table 9: Some resources for transfer learning from high resource languages to growth opportunities

the term, *growth*, over terms such as low resources to refer to languages with more speakers than resources, such as many of the languages in Table 9. Table 9 is sorted by the number of speakers.[12] The columns are based on:

- Articles in Wikipedia[13]
- Joshi classification[14] (Joshi et al., 2020)
- Abstracts in Semantic Scholar (S2)
- Articles in ACL Anthology[15]
- Datasets and Models in HuggingFace (HF)

The good news is that we have more resources these days for growth languages than we had for English when we started EMNLP in 1990s. In addition to the resources in Table 9, there is support for most of these languages in multilingual LLMs, Google Translate, and No Language Left Behind (NLLB) (NLLB Team et al., 2022).

How can we transfer between languages with more resources and languages with fewer resources? The crux of the problem is to construct a comparable corpus of English and the growth language. Given that, there are a number of well-established methods to train language models.

Many efforts start by pivoting from English. That is, they use English documents as the source text, and then translate from English to the growth opportunity. Filter bubbles are a problem for this approach. This approach will not learn aspects of the low resource language that go beyond what is in the high resource language.

We suggest using translation in the reverse direction, as well as similarities based on recommender technologies in academic search engines. That is, we will start with source texts in the growth language such as Wikipedia articles and academic papers in Semantic Scholar (S2) (Wade, 2022). We can then find “nearby” English by several means:

1. Translation from growth language to English
2. Similar in a BERT-like vector space using Specter vectors (Cohan et al., 2020) from S2
3. Similar in terms of random walks on citations

By starting with documents in the growth language, we avoid the filter bubble criticism above. In addition, professional translators specialize in one direction and not the other. They prefer to translate into their stronger language than vice versa. We

[12] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[13] https://en.wikipedia.org/wiki/List_of_Wikipedias

[14] https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt

[15] Based on searches such as https://aclanthology.org/search/?q=hausa

suggest similar logic applies to transfer learning. It is better for systems that are stronger in English to translate into English than vice versa.

3.3 Filter Bubbles: A Monolingual Use Case

3.3.1 Filter Bubbles in News and Academia

There are opportunities for CC to burst filter bubbles, both in monolingual and multilingual applications. With the rise of social media and cable news, we all live in filter bubbles. You may remember EMNLP was in Hong Kong just before COVID. I was interested in the coverage of demonstrations in Hong Kong. The story was very simple in New York and in Beijing. The two perspectives disagreed in many respects, of course, but they agreed on simplicity. When I went to Hong Kong for EMNLP, I learned that the story was anything but simple. In short, we all have a tendency to oversimplify the truth, especially about events that are far way, *of which we know little*,[16] like the famous cover of the New Yorker magazine with a view of the world from 9th avenue.[17]

Ground News has created a business by helping people see their blind spots.[18] They track coverage in a range of different news outlets, and report who is saying what. Is this story covered more by outlets on the left or by outlets on the right?

This is an excellent place to start, but the news is fragmented in many more dimensions than just left/right in America. The conflict in Sryia, for example, overlays three dimensions: (1) America/Russia, (2) Sunni/Shia and (3) Turkey/Kurds. More dimensions are more challenging.

Academic conflicts have even more dimensions. Each school of thought has its position, and its friends and foes. In (Church, 2011), I suggested the pendulum has been swinging back and forth between empiricism and rationalism every 20 years. Here is a slightly updated version of that argument:

- Empiricism I (1950s):
  Shannon, Skinner, Harris, Firth
- Rationalism I (1970s):
  Chomsky, Minsky
- Empiricism II (1990s):
  IBM, AT&T Bell Labs, EMNLP, WWW
- Empiricism III (2010s):
  Deep networks, LLMs, chat bots, RAG

Why is the gap around 20 years? One suggestion involves the cliche that grandparents and grandchildren have a natural alliance. Each academic generation rebels against the their teachers. Chomsky and Minsky rebelled against methods that were popular in the 1950s, and my generation returned the favor by reviving those methods. When we started EMNLP (Empirical Methods in Natural Language Processing), the E-word was an act of rebellion.

3.3.2 How can CC burst these filter bubbles?

Suppose we consider Semantic Scholar to be a CC full of multiple overlays that go well beyond Empiricism and Rationalism. We can model the literature as schools of thought with agreements within clusters and disagreements across clusters.

As suggested above, these days, it has become standard practice to represent everything with vectors. We can use vectors to represent papers, as well as schools of thought. Cosines can be used to estimate agreement and disagreement. There are a number of ways to represent papers as vectors. Two suggestions were mentioned above: BERT-like Specter vectors and random walks on citations[19] (Zhang et al., 2019). We normally use comparable corpora in bilingual applications, but this application, clustering, has applications in both bilingual and monolingual settings.

3.3.3 Comparable Corpora and Bots

Web, news and social media offer many different perspectives and points of view. American bots are trained on American corpora; these bots currently lack a historian's ability to approach conflicts from multiple perspectives.

As homework for my NLP class (Church, 2024), I asked students to write essays about the Opium War from multiple perspectives including both the East and the West. They were encouraged to use bots, but were told they would be responsible for the content. I had hoped students would rewrite output from the bots, but few did. Even students from China handed in essays from an American perspective, because American bots are trained on American corpora. These bots do not mention "the century of humiliation,"[20] a perspective in the East which is motivating efforts to compete with the

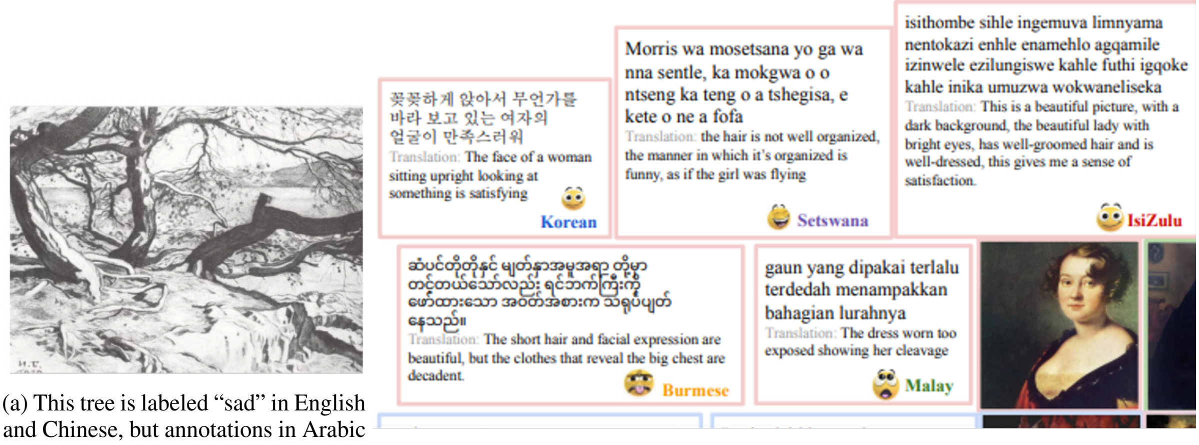
[16] https://www.iwp.edu/articles/2023/04/18/a-quarrel-in-a-faraway-country-between-people-of-whom-we-know-nothing/

[17] https://en.wikipedia.org/wiki/View_of_the_World_from_9th_Avenue

[18] https://ground.news/blindspotter/methodology

[19] https://github.com/VHRanger/nodevectors

[20] https://www.uscc.gov/sites/default/files/3.10.11Kaufman.pdf


꽂꽂하게 앉아서 무언가를 바라 보고 있는 여자의 얼굴이 만족스러워
Translation: The face of a woman sitting upright looking at something is satisfying
Korean
Morris wa mosetsana yo ga wa nna sentle, ka mokgwa o o ntseng ka teng o a tshegisa, e kete o ne a fofa
Translation: the hair is not well organized, the manner in which it's organized is funny, as if the girl was flying
Setswana
isithombe sihle ingemuva limnyama nentokazi enhle enamehlo agqamile izinwele ezilungiswe kahle futhi igqoke kahle inika umuzwa wokwaneliseka
Translation: This is a beautiful picture, with a dark background, the beautiful lady with bright eyes, has well-groomed hair and is well-dressed, this gives me a sense of satisfaction.
IsiZulu
ဆံပင်တိုတိုနှင့် မျက်နှာအမူအရာ တို့မှာ တင့်တယ်သော်လည်း ရင်ဘက်ကြီးကို ဖော်ထားသော အဝတ်အစားက သရုပ်ပျက် နေသည်။
Translation: The short hair and facial expression are beautiful, but the clothes that reveal the big chest are decadent.
Burmese
gaun yang dipakai terlalu terdedah menampakkan bahagian lurahnya
Translation: The dress worn too exposed showing her cleavage
Malay

(a) This tree is labeled "sad" in English and Chinese, but annotations in Arabic are more positive.

(b) Many annotations are positive, but some object to the dress as too revealing.

Figure 1: Emotion labels and captions depend on annotator's background (language/culture).

West in AI so China does not fall behind in technology like it did during the Opium Wars.

Bot technology remains far behind historians like Platt (2019). Bots see the world from a single (American) perspective. Filter bubbles are dangerous; they contribute to trade wars and worse.

3.4 Comparable Corpora and Pictures

We normally think of corpora as text, but now that we are representing everything as vectors, we can generalize corpora to include more modalities: text, speech, pictures, speech, video, etc. As mentioned above, we are worried about pivoting from English prompts. If we start with English prompts, then we are likely to bias responses toward an English perspective. Mohamed et al. (2022, 2024) starts with pictures from WikiArt[21] as prompts. Annotators are asked to add emotion labels and captions in 28 languages, as shown in Figure 1. Different annotators label pictures with different emotion labels and captions, depending on their language and background. The papers refer to a GitHub with a benchmark, as well as baseline implementations of captioning systems that transfer from high resource languages to growth opportunities. Hopefully, the community will accept the challenge and come up with even better systems that embrace diversity over many regions, cultures and languages.

It should be possible to beat a baseline system that translates the captions from English to growth languages. Consider the objections to the dress in Figure 1b. This is a case where it should be possible to outperform a captioning system that translates from English because these objections are unlikely to be found in English captions. In fact, a reviewer asked for an ethics review, objecting to the objections to the dress. We are not siding with one annotator over another, but we object to the objection to the objection. It is not appropriate for us to impose American sensibilities on the rest of the world. Rather than remove biases from corpora (and WikiArt), we hope to build bots that will be more aware of regional sensitivities to topics such as: dress, nudes, religion and alcohol.

4 Conclusions

This paper started with a review of the history of comparable corpora in section 2, followed by a discussion of challenges for the future in section 3.

1. BLI is based on a (too) simple view of the lexicon. Can we capture etymology? Differences between monolingual and bilingual senses?
2. Transfer learning to growth languages: Avoid pivoting via English. Better to prompt with pictures. If we have to translate, it is better to translate into English than out of English to avoid imposing American values on others.
3. Similarities between CC and recommender systems for academic search: can we compare and contrast a query document with candidate recommendations? Can we cluster documents in monolingual and multilingual settings, and compare/contrast within and across clusters?
4. Filter bubbles: chat bots currently lack a historian's ability to approach conflicts from multiple perspectives; bots made in America are trained on corpora from an American perspective with American biases. Can we capture "possible worlds" and diverse perspectives?

[21] https://www.wikiart.org/

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *arXiv preprint arXiv:2304.09919*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Yehoshua Bar-Hillel. 1960. The present status of automatic translation of languages. *Advances in computers*, 1:91–163.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Sebastian Bruch. 2024. *Foundations of Vector Retrieval*. Springer.

Kenneth Church. 2011. A pendulum swung too far. *Linguistic Issues in Language Technology*, 6.

Kenneth Church. 2024. Emerging trends: When can users trust GPT, and when should they intervene? *Natural Language Engineering*, pages 1–11.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *LREC*.

JR Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*.

Pascale Fung. 2000. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. *Parallel Text Processing: Alignment and use of translation corpora*, pages 219–236.

Pascale Fung and Kenneth Ward Church. 1994. K-vec: A new approach for aligning parallel texts. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 414–420, Montreal, Quebec, Canada. Association for Computational Linguistics.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Zellig Harris. 1964. Distributional structure. *Word*, (2):146–162.

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*, pages 6282–6293, Online. Association for Computational Linguistics.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.

Vukosi Marivate, Daniel Njini, Andani Madodonga, Richard Lastrucci, and Jenalea Dzingirai, Isheanesu Rajab. 2023. The vuk'uzenzele south african multilingual corpus.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8770–8785, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, et al. 2024. No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In *EMNLP*, pages 20939–20962, Miami, Florida, USA. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

NLLB Team, Marta Ruiz Costa-jussà, James Cross, Onur Celebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Stephen R Platt. 2019. *Imperial Twilight: The Opium War and the End of China's Last Golden Age*. Vintage.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2023. *Building and Using Comparable Corpora for Multilingual Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Springer Nature.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Alex D Wade. 2022. The semantic scholar academic graph (s2ag). *Companion Proceedings of the Web Conference 2022*.

Jie Zhang, Yuxiao Dong, Yan Wang, Jie Tang, and Ming Ding. 2019. Prone: Fast and scalable network representation learning. In *IJCAI*, volume 19, pages 4278–4284.

SELEXINI – a large and diverse automatically parsed corpus of French

Manon Scholivet[1], Agata Savary[1], Louis Estève[1],
Marie Candito[2], Carlos Ramisch[3],

[1] Université Paris-Saclay, CNRS, LISN, Orsay, France,
[2] Université Paris Cité, CNRS, LLF, Paris, France,
[3] Aix Marseille Univ, CNRS, LIS, Marseille, France

first.last@lisn.fr[1], first.last@u-paris.fr[2], first.last@lis-lab.fr[3]

Abstract

The annotation of large text corpora is essential for many tasks. We present here a large automatically annotated corpus for French. This corpus is divided into two parts: the first from BigScience, and the second from HPLT. The annotated documents from HPLT were selected in order to optimise the lexical diversity of the final corpus SELEXINI. An analysis of the impact of this selection was carried out on syntactic diversity, as well as on the quality of the new words resulting from the HPLT part of SELEXINI. We have shown that despite the introduction of interesting new words, the texts extracted from HPLT are very noisy. Furthermore, increasing lexical diversity did not increase syntactic diversity.

1 Introduction

Morphosyntactic treebanks are cornerstones of grammar induction (Zhu et al., 2020) and of morphosyntactic parsing, whether in monoligual (Dary et al., 2022), multilingual (Straka, 2018) or cross-lingual (Glavaš and Vulić, 2021) contexts. They help to probe language models for linguistic knowledge possibly encoded therein (Shen et al., 2023), and for challenges, e.g. related to syntactic constructions, which these models might fail to appropriately address (Bonial and Tayyar Madabushi, 2024).

Treebanks are also fundamental resources in research on language. They enable studying linguistic properties within or across languages (Levshina et al., 2023), examining the appropriateness of language universals (Brosa-Rodríguez and Kahane, 2024), formalising and searching for complex phenomena such as constructions (Weissweiler et al., 2024a) or documenting low-resourced and endangered languages and dialects (Pugh and Tyers, 2024), inter alia.

For some of such research questions, manually annotated treebanks are not enough to check generalisations and touch upon long-tail phenomena (Sheinfux et al., 2019). In such cases, corpora automatically annotated for morphology (Baroni et al., 2009) and/or syntax (van Noord et al., 2013; Ginter et al., 2013) are used (Schneider, 2011; Bloem et al., 2014).

Our objective is to build such a morphosyntactically parsed corpus for French which would fulfill two conditions. First, it should be large but manageable, i.e. its parsing, storage and maintenance cost should not be prohibitive. Second, it should still have sufficient lexical and syntactic diversity to serve studies in which long-tail phenomena play important roles, such as frame induction (QasemiZadeh et al., 2019), identification of multiword expressions (MWEs) unseen in manually annotated corpora (Ramisch et al., 2020), probing language models for rare but interesting syntactic phenomena (Misra and Mahowald, 2024; Weissweiler et al., 2024b), etc.

To this aim, we use two very large raw corpora: BigScience (Laurençon et al., 2022) and HPLT (**H**igh **P**erformance **L**anguage **T**echnologies) (De Gibert et al., 2024). We select a clean subset of BigScience and we extend it with fragments of HPLT sampled so as to increase the diversity of the whole resulting corpus, henceforth called SELEXINI[1].

Even if both lexical and syntactic diversity are of interest for us, the latter requires pre-existing syntactic annotation, which is prohibitive with a corpus as large as HPLT. Therefore, for data sampling we only use lexical diversity, formally defined as entropy over word types. This sampling strategy likely also has an impact on syntactic diversity, and more generally on the Zipfian distribution of the corpus, as new words and syntactic structures are added and the pre-existing ones change their frequencies. In this context, our research questions

[1] http://hdl.handle.net/11234/1-5822

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 83–98
January 20, 2025. ©2025 Association for Computational Linguistics

are:

- Q1 How does data sampling driven by lexical diversity influence the syntactic diversity of the corpus?
- Q2 What are the resulting quantitative and qualitative properties of the corpus in terms of its Zipfian distribution?

Q1 and Q2 are studied in a comparative context. Namely, we compare BigScience and two extracts of HPLT: one sampled by diversity and another random.

The paper is organized as follows. We briefly discuss related work on French syntactic treebanks (Section 2). We define the diversity measures used for data sampling and corpus comparison (Section 3). We describe the guiding principles (Section 4) used in the corpus construction, as well as the source data (Section 5), their sampling (Section 5.3) and parsing (Section 6). We perform a comparative analysis of two parts of the resulting corpus (Section 7). We finally discuss the limitations of our approach (Section 8) and the conclusions (Section 9).

2 Related work

In dependency syntax, two annotation schemas come with large manually annotated treebanks for French. Historically the FTB-dep schema is a French-specific dependency schema, defined as the result of automatic conversion (Candito et al., 2010) of the 18k phrase-structure trees of the French Treebank (Abeillé et al., 2003). An out-of-domain additional corpus of 3k sentences (the Sequoia corpus (Candito and Seddah, 2012)) is also available in this schema. Then, treebanks of various genres were either annotated under or converted to the Universal Dependencies (UD) schema (Nivre et al., 2020), for a total of 29,735 sentences in UD version 2.15. Concerning available large annotated French corpora, the web-based 1.6 billion token corpus frWac[2] was automatically POS-tagged. Available syntactically parsed corpora are either much smaller (a 150 million token regional news corpus (Seddah et al., 2012)) or mono-genre (parsed French Wikipedia distributed for the CoNLL 2017 shared task[3]).

This shows that no currently existing openly available and morphosyntactically parsed resource is large and diverse enough to serve our needs.

3 Diversity measures

Inspired by formal approaches to diversity (Rényi, 1961; Chao et al., 2014; Morales et al., 2020), we consider it to be a property of populations/systems (here: datasets) whose *elements* can be apportioned into *categories*. For lexical diversity, we define categories as word types and elements as their occurrences in the dataset. For instance, the toy corpus with one sentence from Figure 1(a) contains 8 elements, each one belonging to a different category.

For syntactic diversity, we understand categories as complete syntactic subtrees (where for each node all its children nodes are also included), containing only POS labels and dependency relations. Elements are occurrences of these subtrees in the corpus. Figure 1(b) shows a sample category with two elements in Figure 1(a), highlighted in blue. Figure 1(c) contains another category which does not occur in Figure 1(a), although *y* and *jouent* match the tree fragment in Figure 1(c). This is because the category enrooted in V has to contain all children of V. With this understanding of categories, the example in Figure 1(a) has 5 categories (leaves D, A and PRO, and 2 non-trivial subtrees enrooted in NC and in V) and 8 elements (one per word).

Once elements and categories are defined, diversity can be measures along 3 dimensions: *variety* (which deals with the number of categories), *balance* (which tackles how even the distribution of elements into categories is) and *disparity* (which aggregates pairwise distances between categories). Many diversity measures were proposed in the past, especially in ecology, and most of them are hybrids between at least two of those dimensions. One of them is *richness*, i.e. simply the number of categories n, which is a pure variety. Another one is *entropy* (Shannon and Weaver, 1949), defined by (1), which is a hybrid between variety and balance, where $\Delta_n = \{p_1, ..., p_n\}$ is the distribution of categories. We will use H_{lex} and H_{syn} to refer to entropy over word types and syntactic subtrees, respectively, as defined above.

$$H\left(\Delta_n\right) = -\sum_{i=1}^{n} p_i \log_b\left(p_i\right) \tag{1}$$

In natural language data Zipfian distributions, defined by (2), and their generalisations – Zipf-

[2] https://wacky.sslmit.unibo.it/doku.php?id=corpora

[3] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989

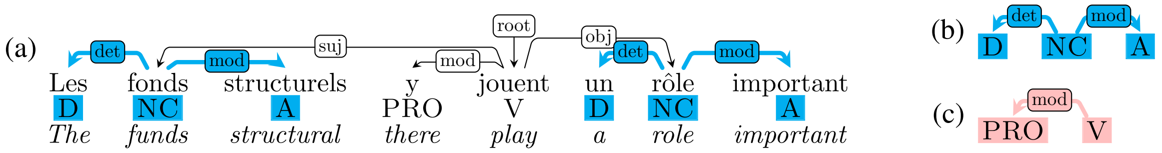

(a)
det
mod
suj
root
mod
obj
det
mod
Les fonds structurels y jouent un rôle important
D NC A PRO V D NC A
The funds structural there play a role important
(b)
det
mod
D NC A
(c)
mod
PRO V

Figure 1: (a) A simplified syntactic tree in FTB-dep schema: *The structural funds play an important role there*; (b) a syntactic category with its two elements highlighted in (a); (c) a syntactic category not occurring in (a), despite the subtree overlap in *y jouent*.

Mandelbrot distributions, defined by (3)[4] – are pervasive. The inverse of their curvature parameter $-s$ can be considered a good balance measure (Lion-Bouton et al., 2022), as it achieves its maximum with $s = 0$, i.e. with a perfectly uniform distribution, and diminishes when the curvature grows (i.e. the data are more and more unbalanced).

$$Z_{s,n}(x) = \left(x^s \sum_{i=1}^{n} i^{-s} \right)^{-1} \tag{2}$$

$$Z^q_{s,n}(x) = (x+q)^{-s} \left(\sum_{i=1}^{n} (i+q)^{-s} \right)^{-1} \tag{3}$$

4 Best practices in corpus construction

There are several best practice recommendations when it comes to creating corpora. The work of (De Pauw, 2006) motivated a number of choices for the construction of this corpus.

Retrieving the data with their **context** makes it possible to analyse the corpus in more detail. It will be easier to understand to whom ‘she’ or ‘he’ refers in a text if we have the text that precedes this sentence. For this, the two corpora on which we are relying (BigScience and HPLT) are ideal because they contain complete documents, which we then segment into sentences.

The data collected must match as closely as possible the language studied (Biber, 1993) in order to obtain a certain level of **representativeness**. This question of representativeness is explored when selecting diversified data. However, the **homogeneity** of the data must not be sacrificed for the sake of diversity. This is why we separate data from the two original corpora (HPLT and BigScience), which are very different in their respective genres (web crawls on the one hand, and parliamentary and Wikipedia texts on the other).

In order for the corpus to be reusable by the community, it is important to use **standards** from the community. Data annotation according to the Universal Dependencies schema was also performed for this reason, in addition to the FTB-dep schema which is specialised for French corpora.

5 Source data

The choice of data to annotate was made in two steps: in the first, preference was given to texts with information on their origins, in order to encourage the use of diversified sources. We focused on French data from the BigScience[5] (Laurençon et al., 2022) as the basis for the SELEXINI corpus. The second selection step was done in order to increase the quantity and the diversity of the corpus data. For this, the HPLT[6] (De Gibert et al., 2024) corpus was chosen. Less clean than BigScience, this part of the corpus nevertheless contains the most diverse part of the data.

5.1 BigScience

The **BigScience** initiative aims to make large quantities of data available in many languages, with the intention of facilitating the training of large multilingual language models (LLMs). Created using pseudo crawls (crawls based on certain predefined domain names), this dataset remains fairly clean.

We chose to work on the parts of the dataset from Europarl, the French part of the United Nations Parallel Corpus and Wikipedia, mainly because of their large size (1.5 billion tokens). Henceforth, this subset will be called BASE. Additionally to its large size, BASE fulfills our other criteria: the metadata allow to easily deduce the language and text genres, no or few multilingual texts are included, licenses are clear and compatible with the intended use of our corpus.[7] The Wikisource subset of BigScience was also considered, but presented too many problems (text starting in the middle of

[4] With $q = 0$ we have $Z_{s,n}(x) = Z^q_{s,n}(x)$.

[5] https://huggingface.co/bigscience-data

[6] https://hplt-project.org/datasets/v1.2

[7] The BigScience RAIL license is inspired both from open licenses and fairness principles: https://bigscience.huggingface.co/blog/the-bigscience-rail-license.

a sentence, HTML tags, encoding problems, sentences in Old French, etc.).

5.2 HPLT

BigScience only contains two text genres: Wikipedia articles and parliamentary debates. In order to achieve a better diversity of genres, we benefit from **HPLT** (De Gibert et al., 2024), a massive multilingual dataset of texts provided by Internet Archive and CommonCrawl. These texts were cleaned by the HPLT authors so as to eliminate documents from dubious URLs (possibly pornographic, racist, etc.) and filter out noisy paragraphs. The remaining documents were then sorted according to he majority vote over a number of language predictors. We work with the cleaned version of French HPLT, containing around 99.59M documents and 122.88B words.

This dataset is still not perfect:

- the filter for setting aside problematic documents is based mainly on the document URL, and some undesired texts can still remain
- the language identification is sometimes erroneous, particularly when several languages are present in the same text
- the data cleaning keeps some uninteresting documents (lists of phone numbers, number plates, etc.)

However, this dataset covers a wide variety of fields and should help increase the diversity of the BASE corpus, as discussed in the following section.

5.3 Diversity-driven data sampling

Diversity of datasets is usually strongly dependent on their sizes. Since we are interested in comparative studies, the compared corpora should have similar sizes. Therefore, we sample HPLT for a subset of a size which would be roughly equivalent to BASE (1.5 billion tokens), while keeping entire texts intact. To reduce computation, we only use a subcorpus containing 6B documents randomly selected from HPLT. We sample it by batches and for each batch we select the document which, added to BASE, maximizes its lexical diversity measured by H in (1). We stop when we exceed the intended size of 1.5 billion tokens. If all batches have been processed and the intended size is not reached, we decrease the size of a batch and reiterate.

The final subset of HPLT selected in this way is called HPLT$_{div}$. For comparison, we also randomly select another subset of HPTL of roughly the same size as HPLT$_{div}$ and we call it HPLT$_{rand}$.

Merging BASE with HPLT$_{div}$ on the one hand, and with HPLT$_{rand}$ on the other hand, yields the final SELEXINI corpus and its non-diverse equivalent SELEXINI$_{rand}$. The following section describes the process of automatic parsing of SELEXINI. Section 7 is then dedicated to comparing the quantitative and qualitative properties of BASE, HPLT$_{div}$ and HPLT$_{rand}$, so as to address the research questions Q1 and Q2.

6 Target annotation schemas and model training

From the outset, we opted for dependency syntax. Morphosyntactic annotation of our corpus can only be done automatically, so to choose the target annotation schemas, we were constrained by the availability of large enough training sets. We thus had two candidates: the monolingual FTB-dep schema or the UD schema (cf. Section 2). We aimed at both accurate linguistic description of French, and cross-lingual parallelism, which exactly corresponds to the balance sought for in the UD project. Yet, for specific linguistic traits, it might prove difficult to satisfy both objectives[8]. Indeed, a closer look at the instantiation of UD guidelines in French UD treebanks first shows some diversity in annotation choices (Guillaume et al., 2019). Second, certain specific phenomena were dealt with (i) either by not following the UD guidelines, which breaks the cross-lingual uniformity, or (ii) by following them at the cost of breaking an internal regularity. We provide some examples in Appendix A.

6.1 Models

For all the previously seen reasons, we chose to keep both annotation schemas, FTB-dep and UD, and thus to build two parsed versions of our SELEXINI corpus, thanks to two models.

FTB-dep To train this model we concatenated two treebanks, containing approximately 21k sentences in total:

- the dependency version of the French Tree-Bank (FTB) (Abeillé et al., 2003), adapted by Seddah et al. (2013);

[8] As put forward in UD's web introduction, which presents UD design as a "subtle compromise" : https://universaldependencies.org/introduction.html.

- the Sequoia[9] treebank (Candito and Seddah, 2012), version 9.2.

While both treebanks have the same main annotation schema (FTB-dep), subtle differences have been introduced over time. In order to get more homogenous training data, we modified the FTB. This harmonisation is described in Appendix B.

UD We use fr_sequoia-ud-2.12 model, one of the models trained on the French treebanks from Universal Dependencies version 2.12, with UDPipe2 (Straka, 2018) and made available by Straka (2023).

6.2 Annotation process

Two different cases were treated to carry out the annotation of the SELEXINI corpus: the annotation with the FTB-dep schema, and the annotation using the Universal Dependencies.

For the annotation using UD, UDPipe 2 was used to carry out all the steps (segmentation, tokenisation, POS tagging, morphological features tagging, lemmas prediction and syntactic analysis).

In the case of the FTB-dep annotation, sentence segmentation and tokenisation were performed using the Bonsai tool[10], designed to specifically handle French. Tagging and parsing were then done with UDPipe 2 as well, but this time using the FTB-dep model described in Section 6.1.

The last step, both for the UD and FTB-dep version, was a lemmas correction phase. While the predicted lemmas on in-domain dev are 99% correct (see Table 1), a qualitative analysis of lemmas for unknown rare word forms on our SELEXINI revealed sometimes absurd predictions[11]. We thus applied lemma correction using the Lefff lexicon (Sagot, 2010)[12].

7 Results

Assessing the quality of annotations is not a trivial task without manually annotated data. We can nevertheless observe the performance of the models on the corresponding dev and test corpora. The results can be seen in Table 1.

Model	Test set	POS	UFeats	LAS
FTB-dep	FTB+Sequoia dev	98.49	94.68	91.11
UD	Sequoia test Gold Tokenisation	99.25	98.01	94.37
	Sequoia test Raw text	98.40	97.19	92.75

Table 1: Scores of the FTB-dep and UD models. The test set for the FTB-dep model is the dev set of the FTB+Sequoia (28 POS tags, 34 dependency labels). For the UD model, the test set of Sequoia is used (17 POS tags, 47 dependency labels)

UDPipe models are frequently used as a baseline thanks to their strong performance. Although the quality of the annotations is better using gold tokenisation than raw text, the results are still good enough to be usable. The model used to annotate in FTB-dep obtains slightly lower scores than the UD model, but as the test corpus and annotation schemes are different, the results are not perfectly comparable and remain acceptable for the annotation task.

We will now compare diversity measures on the different corpora studied. A summary of this information is available in Table 2. The parameters $-s$ et n are computed using equation (3).

7.1 Syntactic Diversity

The algorithm used to select the HPLT$_{div}$ texts aimed to maximize lexical diversity (Section 5.3). We will now evaluate whether this selection also had an impact on syntactic diversity (defined in Section 3) in order to answer our research question Q1.

However, syntactic diversity can only be calculated if we have access to the syntax annotations. The SELEXINI corpus, composed of BASE and HPLT$_{div}$, has been parsed but not HPLT$_{rand}$. Therefore, syntactic diversity is only calculated for the former.

In Table 2, we can observe an increase in the lexical entropy H_{lex} : +0.72 for BASE+HPLT$_{div}$. The opposite trend is visible for syntactic diversity: a decrease of 0.36 point when BASE is augmented with HPLT$_{div}$. Although HPLT$_{div}$ is both more varied (higher n) and more balanced (hieher $-s$) than BASE, which leads to a higher entropy from a lexical point of view (8.10 vs. 7.02), HPLT$_{div}$ is less varied and less balanced than BASE from a

[9] https://deep-sequoia.inria.fr/

[10] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

[11] This is the case e.g. for first person verbal form, rare in the FTB+Sequoia training set. Moreover the predicted lemmas sometimes do not match the predicted POS tag. After lemma correction on out-of-domain 2.64 million tokens, 7000 lemmas were modified using the lexicon heuristic. An analysis of the first 100 corrections revealed only one introduced error, and 99 corrected errors.

[12] The heuristic was to replace any predicted lemma unknown in the lexicon by the longest lemma compatible with this word form and POS tag.

Corpus	Size	n_{lex}	$-s_{lex}$	H_{lex}	n_{syn}	$-s_{syn}$	H_{syn}
BASE	1.54	3.5	-1.250	7.02	167	-1.381	7.17
HPLT$_{div}$	1.56	11.7	-1.182	8.10	124	-1.660	6.25
SELEXINI = BASE + HPLT$_{div}$	3.10	13.6	-1.204	7.74	282	-1.457	6.81
HPLT$_{rand}$	1.56	6.4	-1.138	7.42	-	-	-
SELEXINI$_{rand}$ = BASE + HPLT$_{rand}$	3.10	8.7	-1.187	7.41	-	-	-

Table 2: Summary of the sizes of each corpus in billion tokens, the value of their Zipfian parameters, n for the number of categories in millions (higher is better), and $-s$ for the Zipfian curvature (closer to 0 is better). H is the entropy (higher is better). All these measures are computed for the lexical and syntaxtic version.

syntactic perspective.

As a reminder, n_{lex} and n_{syn} correspond respectively to the number of lexical categories (words) and the number of syntactic categories (syntactic subtrees). The number of common lexical categories between BASE and HPLT$_{div}$ is 1.6 million words, i.e. 11.8% of the total final corpus (BASE + HPLT$_{div}$, i.e. SELEXINI). However, in the case of syntax, there are 9 million common trees, which this time represents only 3.2% of the final corpus.

While 74.3% of the lexical categories in SELEXINI originate from HPLT$_{div}$ only, 41.8% of the syntactic categories originate from HPLT$_{div}$. HPLT$_{div}$ therefore has more weight, more impact, on the diversity of SELEXINI than BASE from a lexical point of view. However, this is not true for syntactic diversity.

Now, if we look at the $-s$ parameter of the Zipfian curvature, which is a measure of balance, we can see that in lexical terms, $-s_{lex}$ obtains a better score for HPLT$_{div}$ than for BASE. This is reversed in the case of s_{syn} where HPLT$_{div}$ is clearly less balanced than BASE.

In conclusion, as an answer to Q1, it appears that optimizing lexical diversity with HPLT$_{div}$ did not also improve syntactic diversity. On the contrary, the sampling had the opposite effect, causing syntactic diversity to notably decrease.

7.2 Lexical Zipfian distributions

In this section, we will focus first on the differences between BASE and HPLT$_{div}$. Then, SELEXINI and PLDH$_{rand}$ will also be compared. In order to answer the research question Q2, we will first carry out an analysis of the quantitative properties by looking at the different scores in Table 2. Secondly, we will analyse the qualitative properties by exploring the new words added by HPLT$_{div}$ to SELEXINI. This section deals only with lexical diversity.

Quantitative properties Although BASE and HPLT$_{rand}$ have roughly the same size, HPLT$_{rand}$ is more diverse than BASE, whether for entropy H_{lex}, variety n_{lex} or balance s_{lex}. One hypothesis is that the Wikipedia articles and parliamentary debates in BigScience create a certain redundancy in the data, making this dataset a less varied and balanced than those from HPLT.

As seen in the previous subsection, augmenting BASE with HPLT$_{div}$ increased the lexical entropy H_{lex} from 7.02 to 7.74: a gain of 0.72. Augmenting BASE with HPLT$_{rand}$ increased the entropy to 7.41: a gain of only 0.4. HPLT$_{rand}$ has roughly half as many categories as HPLT$_{div}$ (11.7 million and 6.4 million respectively). HPLT$_{rand}$ is therefore much less varied than HPLT$_{div}$ (although it is still more varied than BASE). However, with an s_{lex} at 1.182 for HPLT$_{div}$ and at 1.138 for HPLT$_{rand}$, the latter is more balanced. So it is likely that the selection algorithm favours variety more than balance.

Qualitative properties For this section, we extracted the vocabularies of BASE and HPLT$_{div}$. We began by identifying new words present in HPLT$_{div}$ that were not present in BASE. A list of around 10 million words was thus extracted. We then got 2 million static embeddings of dimension 300 from Grave et al. (2018). These embeddings were trained on Common Crawl and Wikipedia using fastText, and keeping only the 2 million most frequent words.We can assume that most words without embeddings will be noise. Only 84,526 of the 10 million words have word embeddings. This means that over 99% of the new words in HPLT$_{div}$ are noise.

Nevertheless, we're going to try to identify whether we can find any common points among the non-noisy words in HPLT$_{div}$. We created word embeddings clusters that can be seen in Figure 2. These clusters were obtained by randomly selecting 2,000 words from our list and extracting their

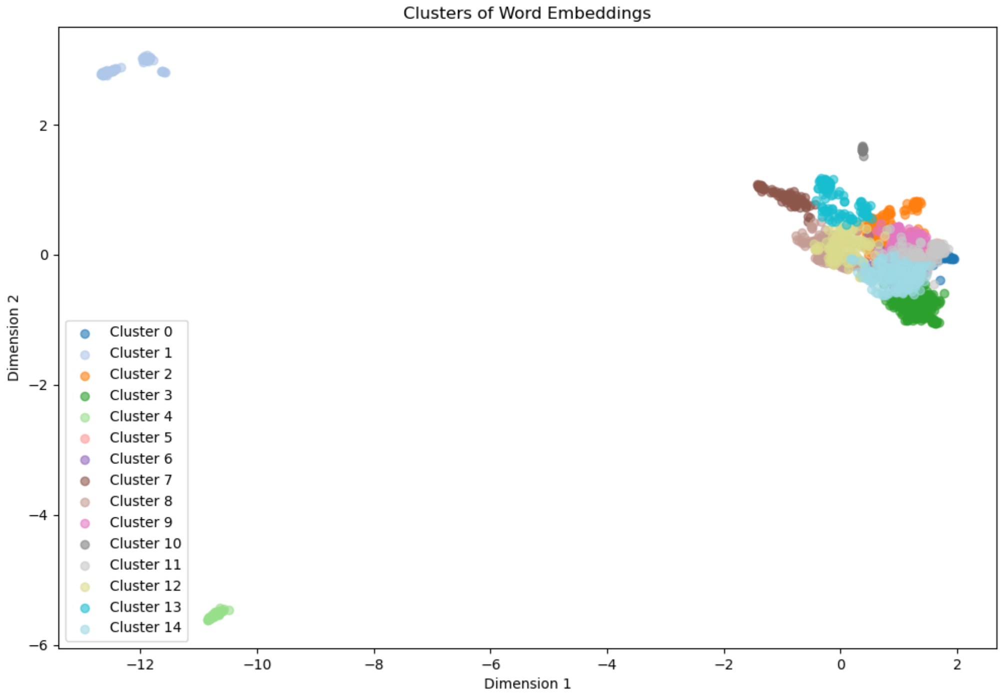


Figure 2: Word embeddings clusters of new words from HPLT$_{div}$.

embeddings. We then reduce the number of dimensions to 50 using the UMAP algorithm (McInnes et al., 2018). We cluster our embeddings using the K-means algorithm, with K=15[13]. Finally, we reduce our embeddings again to two dimensions using the PCA algorithm, in order to visualise the clusters. Details of the clusters are available in Appendix C. Although the clusters are not perfectly pure, common themes can be identified across most of them.

- Clusters 1 and 4, very isolated from the rest, contain only dates in two different formats (year-month-day for cluster 1 and day-month-year for cluster 4). Cluster 10 also contains numbers only, with a '-'.
- Cluster 8 contains 'sms'-type language: *copinette* (friend), *choupie* (cute) or *merciiii* (thanks).
- Clusters 2, 7 and 13 contain words in other languages : *bedsheets* (English), *polozoni* (Croatian) or *abgerufen* (German).
- Cluster 2 also contains neologisms and concatenations of words: *brocantitude* (flea market attitude), *miseenservice* (commissioning)
- Cluster 7 contains a subcluster with symbols and emojis
- Clusters 0, 5, 11 and 14 contains many spelling mistakes, often due to missing accents : *patrimoin* (heritage), *helices* (propeller), *ludotheque* (toy library), *mesage* (message)
- Clusters 5 and 11 also contain suffixes : *ficiaires*, *ctions*, *geait*, *pondants*
- Cluster 3 contains rare forms of conjugation : *flippent* (they freak out), *débuterez* (you will start) or *chouchoutent* (they pamper)
- Cluster 6 contains words concatenated with a final dot : *normalement.* (normally.), *châteaux.* (castles.), *surf.* (surf.)
- Cluster 12 contains URLs and filenames : *main.php*, *monsite.com*, *top-site*
- Cluster 9 is the only cluster with no specific theme. There are misspells (*pâtissiére* (fe-

[13] The choice of 15 clusters was made empirically.

male baker)), rare words (*non-couvert* (not covered)), foreign word (*cocoon*) and others

In conclusion, although the quantitative study showed that texts that increase variety are more often selected (either because of their greater importance in the entropy, or because they occur more frequently than texts that increase the balance), the qualitative study showed that this variety is almost artificial, because of the very high noise content of the texts from HPLT. Nevertheless, some new "valid" word forms are added, especially rare conjugations, which usefully extend the vocabulary of SELEXINI.

8 Limitations and future work

A first limitation of this work is obviously the presence of a lot of noise in HPLT. Applying the selection algorithm to a corpus without noise could lead to very different results and conclusions. The use of noise reduction techniques could also help to limit the problem (Zhu et al., 2022).

Another limitation is the automatic prediction of labels. These predictions carry the biases of the models used to generate these annotations, which may have only encountered certain rare phenomena on an infrequent basis.

There are many different measures of diversity. Here we focused only on Zipfian parameters and Shannon-Weaver entropies, but some other measures highlight other information. In particular, disparity is another dimension of diversity that we have not explored here, but which would have its rightful place in an analysis of corpus diversity.

9 Conclusions

In this article, we have presented three contributions. The first is the creation of a large automatically parsed French corpus. The second is a study of the impact of lexical diversity-driven data sampling on syntactic diversity. Finally, we also performed a quantitative and qualitative analysis of the lexical diversity resulting from the selection aimed at maximising this same lexical diversity.

The main conclusions are that the selection based on lexical diversity favours variety more than balance, and mainly extracts noise. We also found that there was no positive impact on syntactic diversity, and even that there was a rather negative impact. It would be interesting to understand if this negative impact is due to noisy data or if it is inherent to natural language (e.g. rare and new words might tend to occur in syntactic constructions known for frequent words). More research is still needed to find methods that will maximise lexical diversity while avoiding the problems of noisy texts.

References

Anne Abeillé and Nicolas Barrier. 2004. Enriching a French treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. *Building a Treebank for French*, pages 165–187. Springer Netherlands, Dordrecht.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*.

Douglas Biber. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257.

Jelke Bloem, Arjen Versloot, and Fred Weerman. 2014. Applying automatically parsed corpora to the study of language variation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1974–1984, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Claire Bonial and Harish Tayyar Madabushi. 2024. A construction grammar corpus of varying schematicity: A dataset for the evaluation of abstractions in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 243–255, Torino, Italia. ELRA and ICCL.

Antoni Brosa-Rodríguez and Sylvain Kahane. 2024. New proposal of greenberg's universal 14 from typometrics. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12217–12226, Torino, Italia. ELRA and ICCL.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Ricardo Cordeiro. 2020. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2).

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical French dependency parsing: Treebank conversion and first results. In *Proceedings of the*

Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Marie Candito and Djamé Seddah. 2012. Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France.

Anne Chao, Chun-Huo Chiu, and Lou Jost. 2014. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45:297–324. Publisher: Annual Reviews.

Franck Dary, Maxime Petit, and Alexis Nasr. 2022. Dependency parsing with backtracking using deep reinforcement learning. *Transactions of the Association for Computational Linguistics*, 10:888–903.

Ona De Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer Van Der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, et al. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128.

Guy De Pauw. 2006. Developing Linguistic Corpora—A Guide to Good PracticeMartin Wynne (ed.). *Literary and Linguistic Computing*, 22(1):101–102.

Filip Ginter, Jenna Nyblom, Veronika Laippala, Samuel Kohonen, Katri Haverinen, Simo Vihjanen, and Tapio Salakoski. 2013. Building a large automatically parsed corpus of Finnish. In *Proceedings of the 19th Nordic Conference on Computational Linguistics (NoDaLiDa'13)*, pages 291–300.

Goran Glavaš and Ivan Vulić. 2021. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888, Online. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2022. The bigscience roots corpus: A 1.6tb composite multilingual dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. *Preprint*, arXiv:2403.19827.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphael Fournier-S'niehotta, Remy Poulain, Lionel Tabourier, and Fabien Tarissan. 2020. Measuring Diversity in Heterogeneous Information Networks. *arXiv preprint*. Issue: arXiv:2001.01296 arXiv:2001.01296 [cs, math].

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.

Robert Pugh and Francis Tyers. 2024. A Universal Dependencies treebank for Highland Puebla Nahuatl.

In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1393–1403, Mexico City, Mexico. Association for Computational Linguistics.

Behrang QasemiZadeh, Miriam R. L. Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30, Minneapolis, Minnesota, 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.

Alfréd Rényi. 1961. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4.1, pages 547–562. University of California Press.

Benoît Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Gerold Schneider. 2011. Using automatically parsed corpora to discover lexico-grammatical features of english varieties. In *30th International Conference on Lexis and Grammar, Nicosia, Cyprus*.

Djamé Seddah, Marie Candito, Benoit Crabbé, and Enrique Henestroza Anguiano. 2012. Ubiquitous usage of a broad coverage French corpus: Processing the Est Republicain corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3249–3254, Istanbul, Turkey. European Language Resources Association (ELRA).

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Wolińsk, Alina Wróblewska, and Éric Villemonte de La Clergerie. 2013. Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, United States. Association for Computational Linguistics.

Claude Elwood Shannon and Warren Weaver. 1949. *A Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Livnat Herzig Sheinfux, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner. 2019. Verbal multiword expressions: Idiomaticity and flexibility. In Yannick Parmentier and Jakub Waszczuk, editors, *Representation and parsing of multiword expressions: Current trends*, pages 35–68. Language Science Press, Berlin.

Gaofei Shen, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2023. Wave to syntax: Probing spoken language models for syntax. In *Proc. INTERSPEECH 2023*, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 1259–1263. Publisher Copyright: © 2023 International Speech Communication Association. All rights reserved.

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.

Milan Straka. 2023. Universal dependencies 2.12 models for UDPipe 2 (2023-07-17). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. *Large Scale Syntactic Annotation of Written Dutch: Lassy*, pages 147–164. Springer Berlin Heidelberg, Berlin, Heidelberg.

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. 2024a. UCxn: Typologically informed annotation of constructions atop Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16919–16932, Torino, Italia. ELRA and ICCL.

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2024b. Hybrid human-llm corpus construction and llm evaluation for rare linguistic phenomena. *Preprint*, arXiv:2403.06965.

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Adelani, and Dietrich Klakow. 2022. Is BERT robust to label noise? a study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 62–67, Dublin, Ireland. Association for Computational Linguistics.

Hao Zhu, Yonatan Bisk, and Graham Neubig. 2020. The return of lexical dependencies: Neural lexicalized PCFGs. *Transactions of the Association for Computational Linguistics*, 8:647–661.

A Examples of difficulties in the UD annotation of French

In the French UD treebanks, certain specific phenomena were dealt with (i) either by not following the UD guidelines, which breaks the cross-lingual uniformity, or (ii) by following them at the cost of breaking an internal regularity. As examples of (i), (Guillaume et al., 2019) explicitely report not to follow (for now) UD guidelines for copula constructions with clausal predicative complements (which would lead to a verb with two distinct subjects), nor for expletive *il* subjects[14]. An example of (ii) is the use of different dependency labels for dependents of verb, depending on the category of the dependent, differently to what occurs in the FTBdep annotation schema, itself deriving from the FTB annotation (Abeillé and Barrier, 2004). For instance, the verb *souhaiter (to wish)* can take , the direct a direct complement which is either a NP, an infinitival clause, a clause, or a clitic pronoun. All these cases fill the same valency slot (and thus are mutually exclusive) and are pronominalized using the same accusative clitic pronoun *le*. This uniformity is captured by using a single `obj` label in FTBdep, but 3 different labels in UD (`obj`, `xcomp`, `ccomp`). Moreover, the two latter labels are also used for indirect complements, which obfuscates the linking to semantic roles. Another example concerns the use of `iobj`. For instance for *X parle de Y à Z (X talks about Y to Z)*, in UD, the Y argument can be `iobj`, `obl:arg`, `xcomp`, and the Z argument can be `iobj` or `obl:arg`, whereas Y and Z are uniformly annotated as `de_obj` and `a_obj` in FTBdep.

B FTB modifications

The FTB has been modified compared to the version described in (Seddah et al., 2013). Corrections were done:

- Automatic corrections to ensure flat representations of MWEs have their linearly first component as head of all other components;
- Manual removal of spurious cycles in surface dependency trees (10 cases).

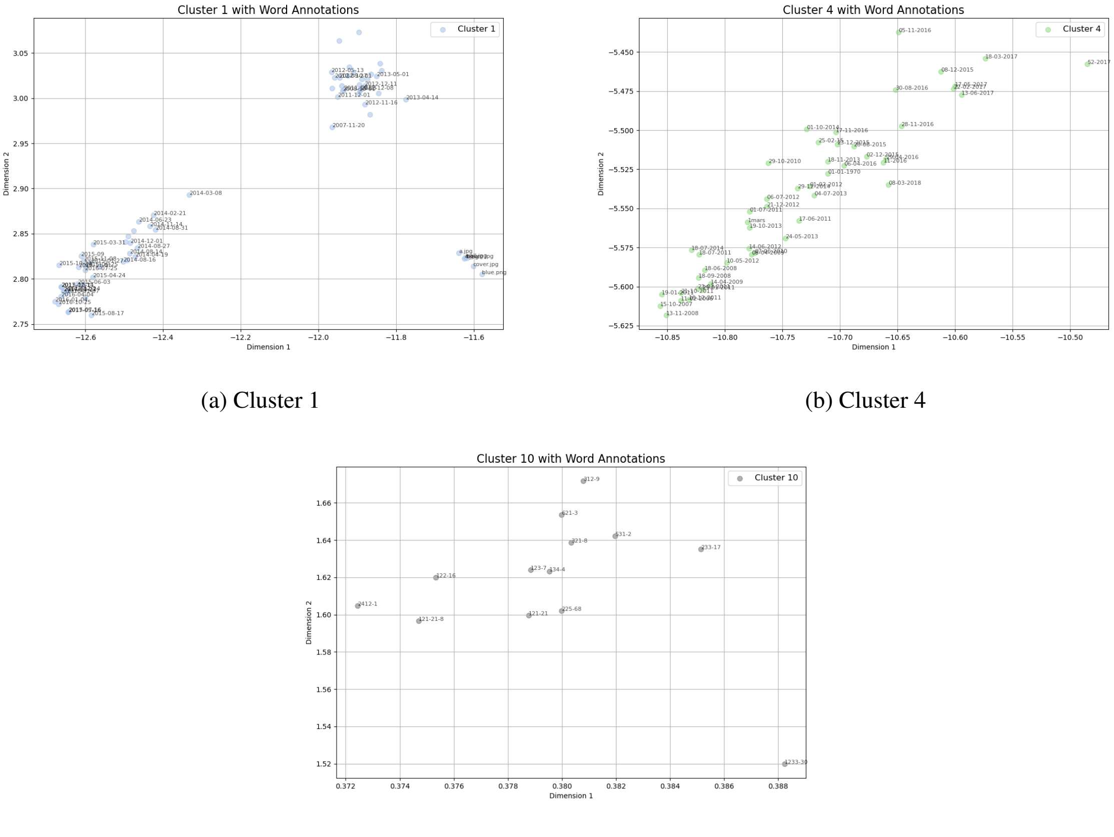
Some harmonisation with the Sequoia treebank :

- Representation of MWEs as in Sequoia 9.2, namely as designed in the PARSEME-FR project[15] and described in (Candito et al., 2020);
  - the main change concerns using regular syntax for MWEs whenever possible;
  - for remaining MWEs, final prepositions or complementizers are not included in the MWE (i.e. *que (that)* not included in the MWE *étant donné (given)*.)
- Minor modifications of tokenization:
  - any X - X (- X)* sequence of tokens within a MWE was remerged as one token (i.e. "au - dessus de" → "au-dessus de")
  - numbers: any sequence [0-9]+ (, [0-9]+)+ merged as one token (i.e. "34 , 7" => "34,7")
- Homogenisation of lemmas:
  - reflexive clitics (CLR tag) have lemma *se*;
  - dative and accusative first and second person clitics all receive *le/lui* lemma (ambiguity is to be solved in syntax);
  - distinguish lemma for *madame (madam)* from that of *monsieur (mister)*.

C Word embeddings clusters

[14] UD guidelines take into account the semantic property of not baring a semantic role, which has clear advantages for downstream semantic analysis, but which causes peculiarities from the stricter syntactic point of view.

[15] `https://parsemefr.lis-lab.fr`



(a) Cluster 1

(b) Cluster 4

(c) Cluster 10

Figure 3: Word embeddings clusters of new words from HPLT$_{div}$ with dates

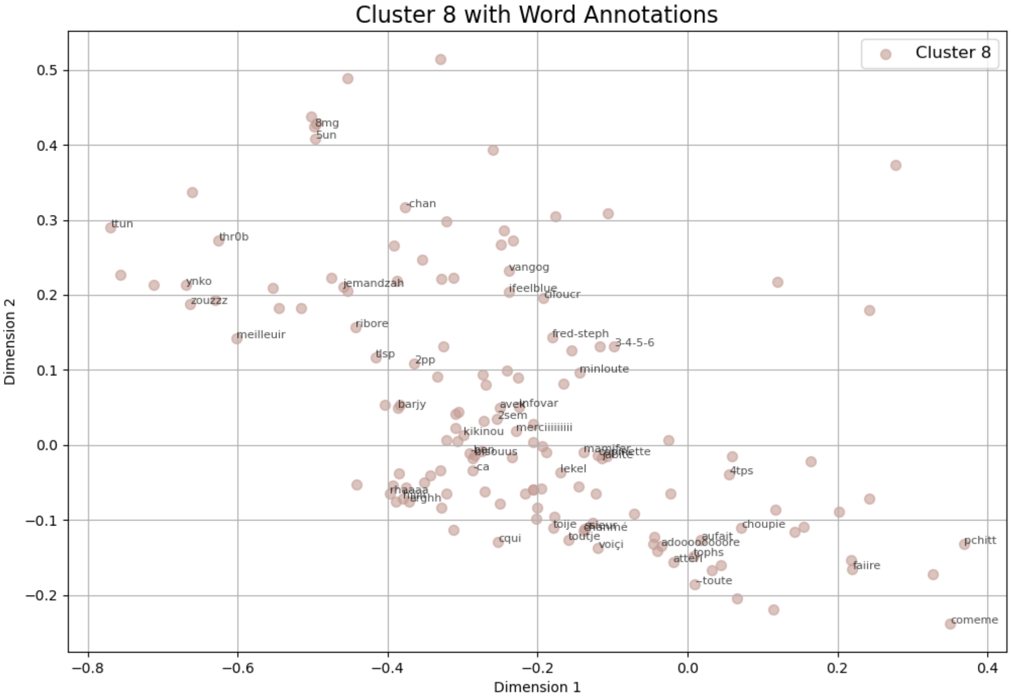
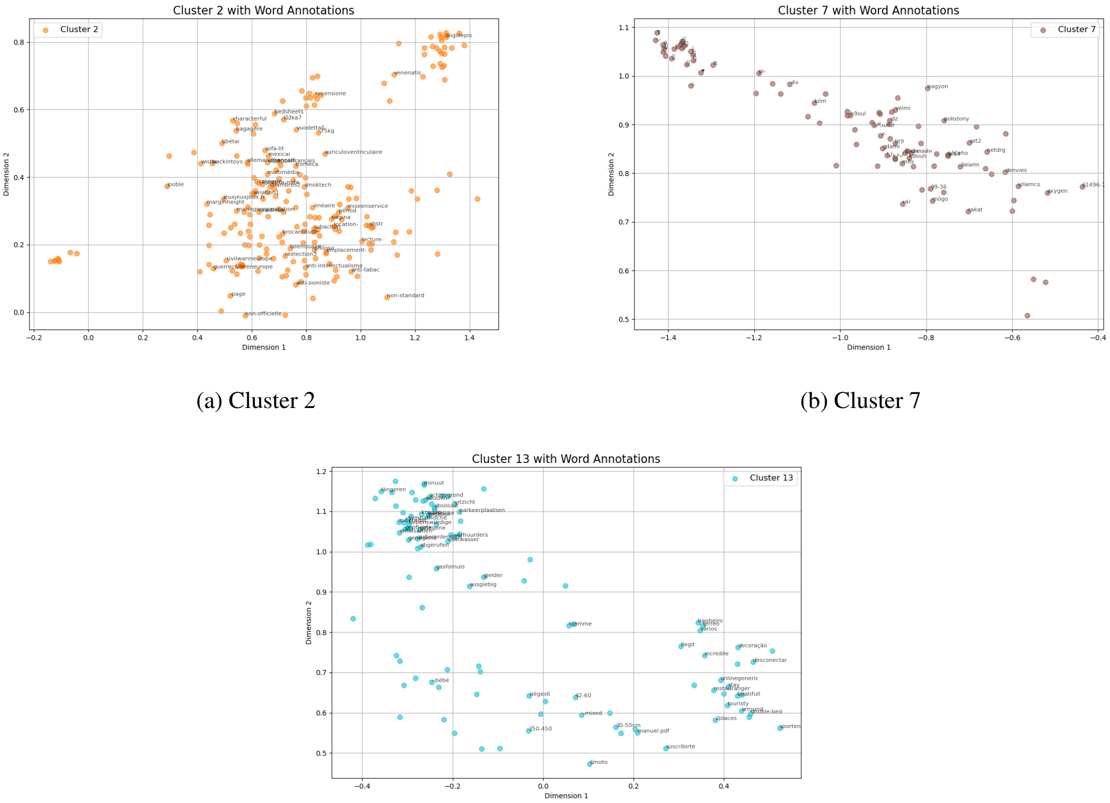


Figure 4: Cluster 8 (SMS language) of Word embeddings clusters of new words from HPLT$_{div}$.



(a) Cluster 2

(b) Cluster 7

(c) Cluster 13

Figure 5: Word embeddings clusters of new words from HPLT$_{div}$ with foreign words

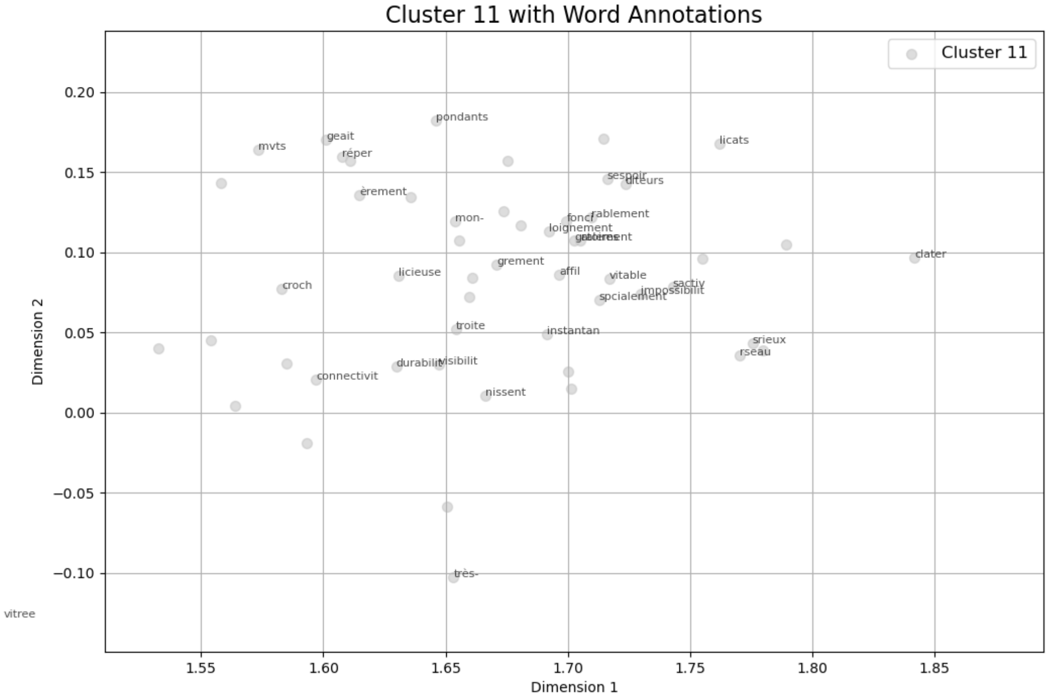
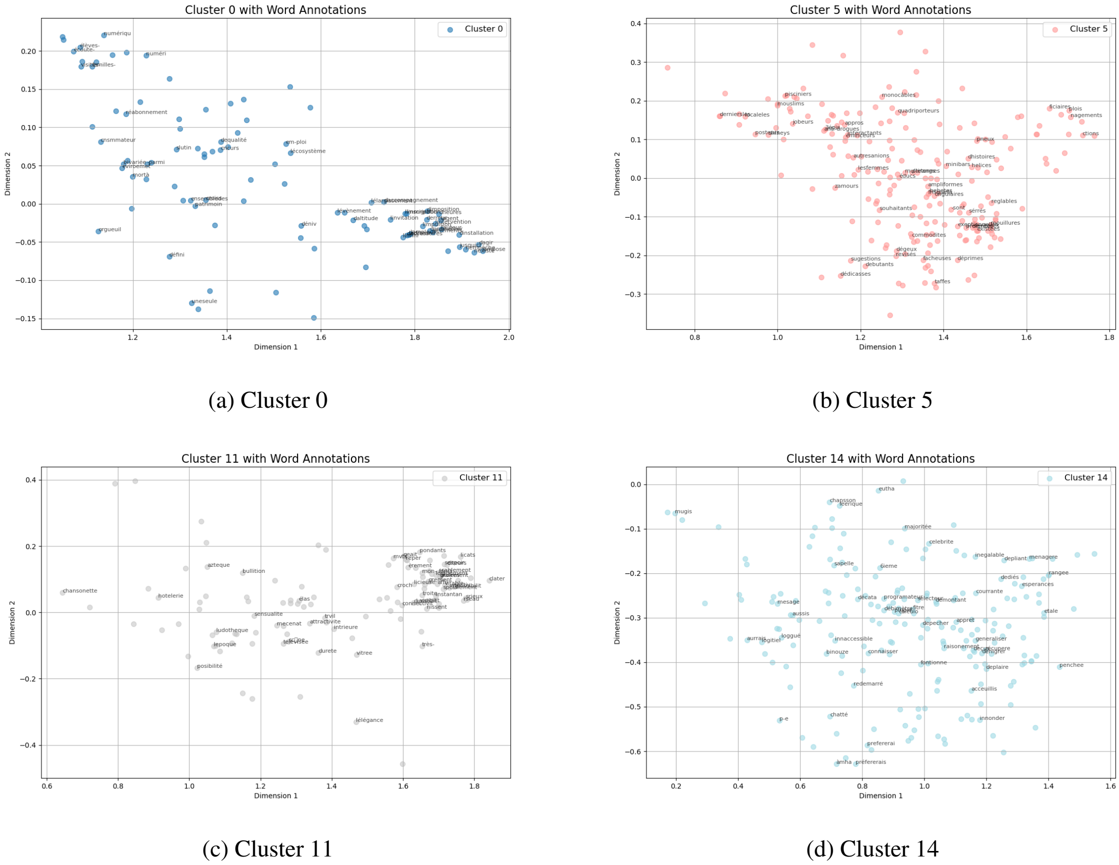


Figure 6: Cluster 11 (suffixes) of Word embeddings clusters of new words from HPLT$_{div}$ (zoom on suffixes part).



(a) Cluster 0

(b) Cluster 5

(c) Cluster 11

(d) Cluster 14

Figure 7: Word embeddings clusters of new words from HPLT$_{div}$ with spelling mistakes

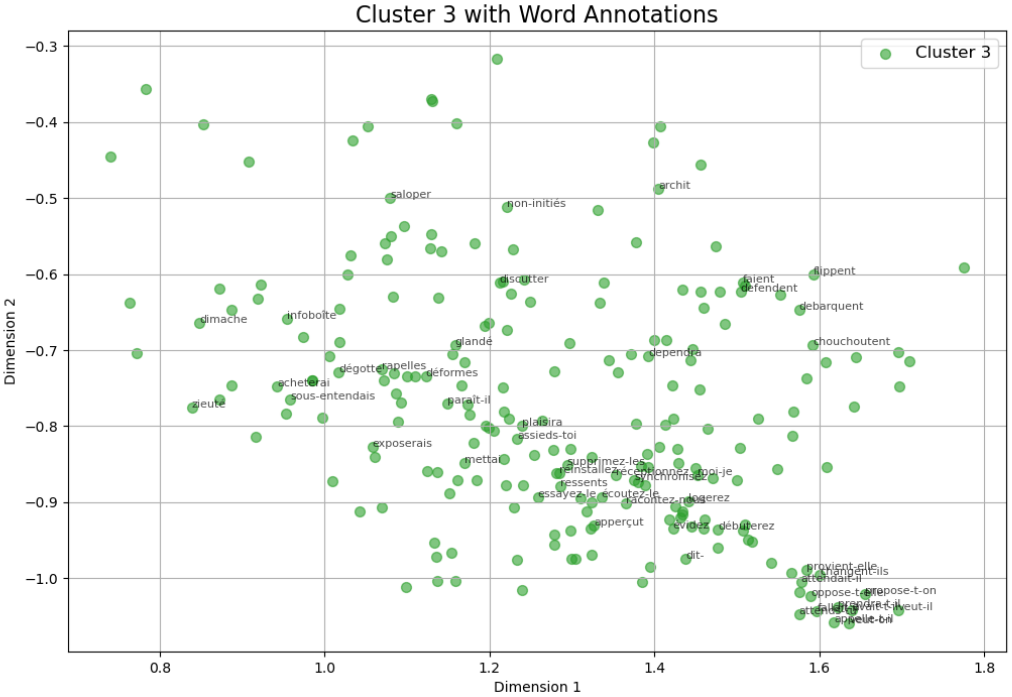


Figure 8: Cluster 3 (rare conjugations) of Word embeddings clusters of new words from HPLT$_{div}$.

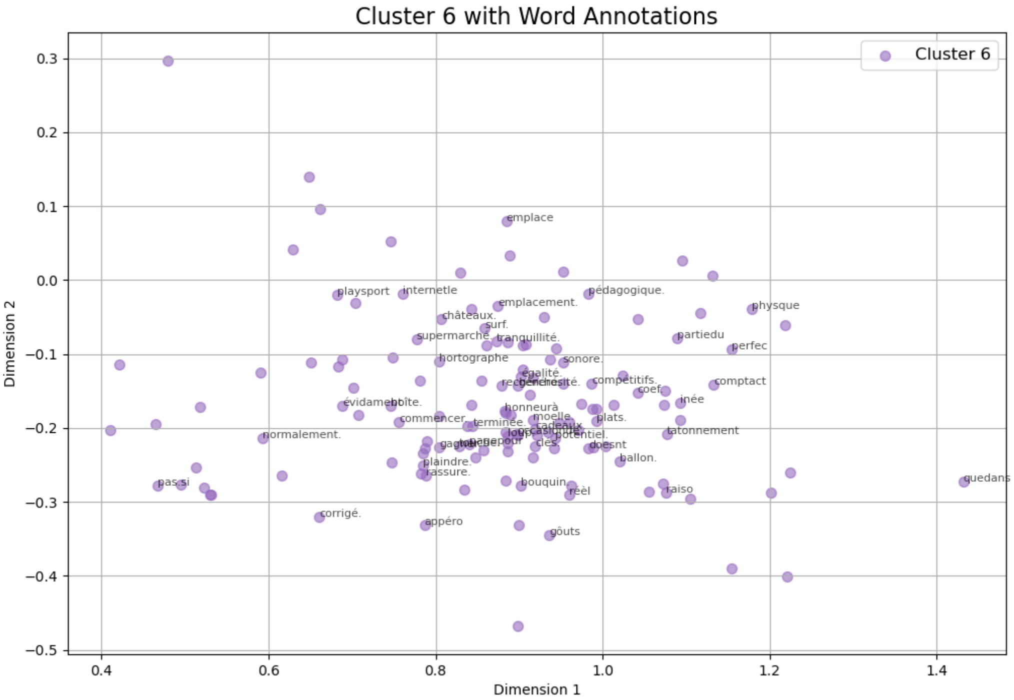


Figure 9: Cluster 6 (final point) of Word embeddings clusters of new words from HPLT$_{div}$.

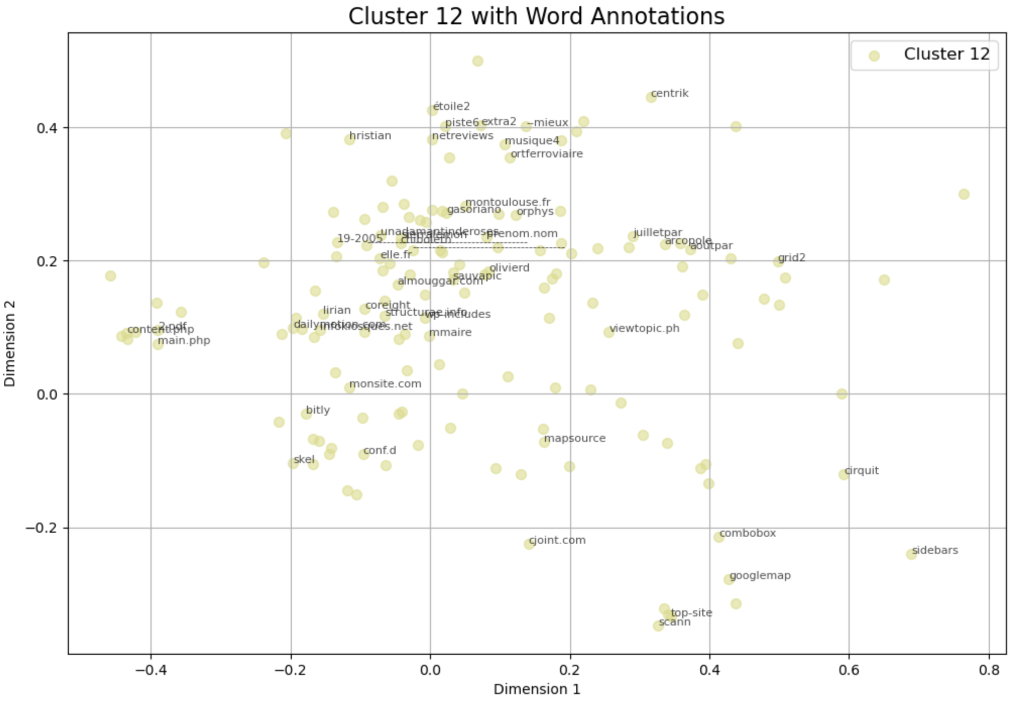


Figure 10: Cluster 12 (url and filenames) of Word embeddings clusters of new words from HPLT$_{div}$.

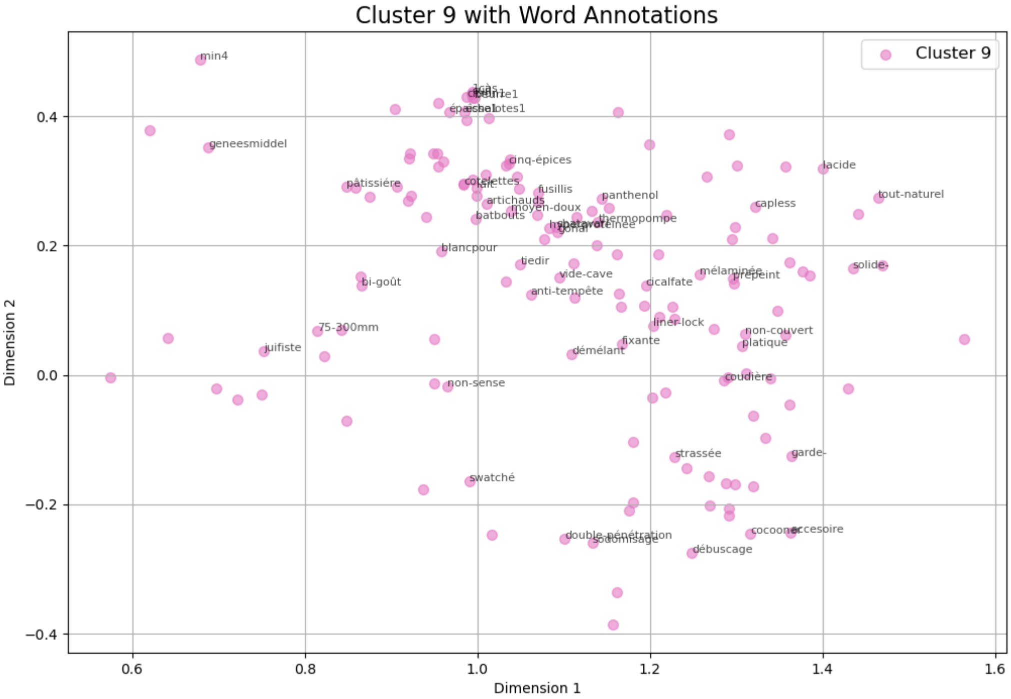


Figure 11: Cluster 9 (diverse) of Word embeddings clusters of new words from HPLT$_{div}$.

Author Index