

LREC-COLING 2024

**ISA-20: The 20th Joint ACL - ISO Workshop
on Interoperable Semantic Annotation
@LREC-COLING-2024**

Workshop Proceedings

Editor
Harry Bunt

20 May, 2024
Torino, Italia

Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @LREC-COLING-2024

Copyright ELRA Language Resources Association (ELRA), 2024
These proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-32-6
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

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

Preface

Twenty years after the first ISA workshop (Tilburg, 2004), this is the year of the 20th edition of the series. The ISA workshops were inspired by the decision of the International Organisation for Standards ISO to start developing annotation standards for language data, including lexical information, typed feature structures, morphological and syntactic information, and the Semantic Annotation Framework (SemAF), a multi-part standard for annotating aspects of meaning. As the development of such standards is the work of small groups of experts nominated by ISO member countries, the ISA workshops were set up with the intention to (a) promote the involvement of all interested scholars in these processes, and (b) to inform scholars in language studies and developers of language resources and linguistic applications of the ISO activities and the standards under development. To support this two-way interaction, the ISA workshops were organised as a joint initiative of the ACL Special Interest Group in Semantics (SIGSEM) and of ISO Working Group TC 37/SC 4/WG 2, Semantic annotation. apart from ISA-3 (Marina del Rey, 2008) and ISA-8 (Pisa, 2012), all ISA workshops have been organised as satellite events of large conferences such as LREC, ACL, IWCS and COLING.

This year's workshop at LREC-COLING 2024 has seen a higher number of submissions, full papers as well as short papers, than any previous edition of the ISA series. Since most of the submissions were of excellent quality, we have been forced to follow the LREC policy of having not only the accepted short papers but also some of the accepted full papers presented as a poster plus a flash presentation. True to the original intention of the ISA workshop series, the ISA-20 program features a mix of papers, presented in these proceedings, on developing new ISO standards such as VoxML (visual information) and QuantML (quantification), on ways to use multiple standards defined by SemAF parts, and on topics not directly related to ISO activities but to semantic annotation as such.

Many thanks are due to the Program Committee members for their diligent and fast review work.

The organisers,

Harry Bunt, Nancy Ide, Kiyong Lee, Volha Petukhova, James Pustejovsky, and Laurent Romary.

Organizing Committee

Harry Bunt
Nancy Ide
Kiyong Lee
Volha Petukhova
James Pustejovsky
Laurent Romary

Program Committee

Jan Alexandersson
Maxime Amblard
Johan Bos
Harry Bunt
Jae-Woong Choe
Stergios Chatzykiriakidis
Robin Cooper
Rodolfo Delmonte
David DeVault
Simon Dobnik
Jens Edlund
Alex Fang
Robert Gaizauskas
Koiti Hasida
Nancy Ide
Elisabetta Jezek
Kiyong Lee
Philippe Muller
Rainer Osswald
Catherine Pelachaud
Volha Petukhova
Laurent Prévot
Stephen Pulman
James Pustejovsky
Laurent Romary
Purificação Silvano
Manfred Stede
Thorsten Trippel
Carl Vogel
Menno van Zaanen
Annie Zaanen
Heike Zinsmeister

Table of Contents

| | |
|---|-----|
| <i>The MEET Corpus: Collocated, Distant and Hybrid Three-party Meetings with a Ranking Task</i> Ghazaleh Esfandiari-Baiat and Jens Edlund | 1 |
| <i>MSNER: A Multilingual Speech Dataset for Named Entity Recognition</i> Quentin Meeus, Marie-Francine Moens and Hugo Van hamme | 8 |
| <i>Attitudes in Diplomatic Speeches: Introducing the CoDipA UNSC 1.0</i> Mariia Anisimova and Šárka Zikánová | 17 |
| <i>Automatic Alignment of Discourse Relations of Different Discourse Annotation Frameworks</i> Yingxue Fu | 27 |
| <i>A New Annotation Scheme for the Semantics of Taste</i> Teresa Paccosi and Sara Tonelli | 39 |
| <i>What to Annotate: Retrieving Lexical Markers of Conspiracy Discourse from an Italian-English Corpus of Telegram Data</i> Costanza Marini and Elisabetta Jezek | 47 |
| <i>Lightweight Connective Detection Using Gradient Boosting</i> Mustafa Erolcan Er, Murathan Kurfalı and Deniz Zeyrek | 53 |
| <i>Shallow Discourse Parsing on Twitter Conversations</i> Berfin Aktas and Burak Özmen | 60 |
| <i>Search tool for An Event-Type Ontology</i> Nataliia Petliak, Cristina Fernández Alcaina, Eva Fučíková, Jan Hajič and Zdeňka Urešová 66 | |
| <i>Tiny But Mighty: A Crowdsourced Benchmark Dataset for Triple Extraction from Unstructured Text</i> Muhammad Salman, Armin Haller, Sergio J. Rodriguez Mendez and Usman Naseem . | 71 |
| <i>Less is Enough: Less-Resourced Multilingual AMR Parsing</i> Bram Vanroy and Tim Van de Cruys | 82 |
| <i>MoCCA: A Model of Comparative Concepts for Aligning Constructicons</i> Arthur Lorenzi, Peter Ljunglöf, Ben Lyngfelt, Tiago Timponi Torrent, William Croft, Alexander Ziem, Nina Böbel, Linnéa Bäckström, Peter Uhrig and Ely E. Matos | 93 |
| <i>ISO 24617-8 Applied: Insights from Multilingual Discourse Relations Annotation in English, Polish, and Portuguese</i> Aleksandra Tomaszewska, Purificação Silvano, António Leal and Evelin Amorim | 99 |
| <i>Combining semantic annotation schemes through interlinking</i> Harry Bunt | 111 |
| <i>Fusing ISO 24617-2 Dialogue Acts and Application-Specific Semantic Content Annotations</i> Andrei Malchanau, Volha Petukhova and Harry Bunt | 122 |

| | |
|---|-----|
| <i>Annotation-Based Semantics for Dialogues in the Vox World</i> Kiyong Lee | 133 |
| <i>Annotating Evaluative Language: Challenges and Solutions in Applying Appraisal Theory</i> Jiamei Zeng, Min Dong and Alex Chengyu Fang | 144 |
| <i>Attractive Multimodal Instructions, Describing Easy and Engaging Recipe Blogs</i> Ielka van der Sluis and Jarred Kiewiet de Jonge | 152 |

Workshop Program

09:00 **Opening; Session 1**

The MEET Corpus: Collocated, Distant and Hybrid Three-party Meetings with a Ranking Task

Ghazaleh Esfandiari-Baiat and Jens Edlund

MSNER: A Multilingual Speech Dataset for Named Entity Recognition

Quentin Meeus, Marie-Francine Moens and Hugo Van hamme

Attitudes in Diplomatic Speeches: Introducing the CoDipA UNSC 1.0

Mariia Anisimova and Šárka Zikánová

10:30 **Coffee break**

11:00 **Session 2**

Automatic Alignment of Discourse Relations of Different Discourse Annotation Frameworks

Yingxue Fu

11:30 **Session 3: Flash presentations**

A New Annotation Scheme for the Semantics of Taste

Teresa Paccosi and Sara Tonelli

What to Annotate: Retrieving Lexical Markers of Conspiracy Discourse from an Italian-English Corpus of Telegram Data

Costanza Marini and Elisabetta Jezek

Lightweight Connective Detection Using Gradient Boosting

Mustafa Erolcan Er, Murathan Kurfalı and Deniz Zeyrek

Shallow Discourse Parsing on Twitter Conversations

Berfin Aktas and Burak Özmen

No Day Set (continued)

Search tool for An Event-Type Ontology

Nataliia Petliak, Cristina Fernández Alcaína, Eva Fučíková, Jan Hajič and Zdeňka Urešová

Tiny But Mighty: A Crowdsourced Benchmark Dataset for Triple Extraction from Unstructured Text

Muhammad Salman, Armin Haller, Sergio J. Rodríguez Mendez and Usman Naseem

Less is Enough: Less-Resourced Multilingual AMR Parsing

Bram Vanroy and Tim Van de Cruys

MoCCA: A Model of Comparative Concepts for Aligning Constructicons

Arthur Lorenzi, Peter Ljunglöf, Ben Lyngfelt, Tiago Timponi Torrent, William Croft, Alexander Ziem, Nina Böbel, Linnéa Bäckström, Peter Uhrig and Ely E. Matos

12:15 **Poster visits**

13:00 **Lunch break**

14:00 **Session 4**

ISO 24617-8 Applied: Insights from Multilingual Discourse Relations Annotation in English, Polish, and Portuguese

Aleksandra Tomaszewska, Purificação Silvano, António Leal and Evelin Amorim

Combining semantic annotation schemes through interlinking

Harry Bunt

Fusing ISO 24617-2 Dialogue Acts and Application-Specific Semantic Content Annotations

Andrei Malchanau, Volha Petukhova and Harry Bunt

Annotation-Based Semantics for Dialogues in the Vox World

Kiyong Lee

No Day Set (continued)

16:00 **Tea break**

16:30 **Session 5**

Annotating Evaluative Language: Challenges and Solutions in Applying Appraisal Theory

Jiamei Zeng, Min Dong and Alex Chengyu Fang

Attractive Multimodal Instructions, Describing Easy and Engaging Recipe Blogs

Ielka van der Sluis and Jarred Kiewiet de Jonge

17:30 **Closing**

The MEET Corpus: Collocated, Distant and Hybrid Three-party Meetings with a Ranking Task

Ghazaleh Esfandiari-Baiat, Jens Edlund

Speech, Music and Hearing, KTH

Stockholm, Sweden

geb@kth.se, edlund@speech.kth.se

Abstract

We introduce the MEET corpus. The corpus was collected with the aim of systematically studying the effects of collocated (physical), remote (digital) and hybrid work meetings on collaborative decision-making. It consists of 10 sessions, where each session contains three recordings: a collocated, a remote and a hybrid meeting between three participants. The participants are working on a different survival ranking task during each meeting. The duration of each meeting ranges from 10 to 18 minutes, resulting in 380 minutes of conversation altogether. We also present the annotation scheme designed specifically to target our research questions. The recordings are currently being transcribed and annotated in accordance with this scheme.

Keywords: meetings, multimodal corpora, annotation scheme

1. Introduction

The declaration of COVID-19 as a global pandemic led to widespread implementation of social distancing measures, resulting in a shift of various human social activities from offline to online. In other words, the enforced social isolations in the physical world significantly increased humans' social interactions in the cyber world (Yan, 2020). In a professional context, this was most noticeable as a shift from collocated to remote meetings that changed the work environment. The shift is backed by staggering numbers. As an example, Zoom added >2 million active users monthly during 2020 (Video Conferencing Market Size, Share & Covid-19 Impact Analysis, 2021).

The change subsists in the post-pandemic era, and remote meetings have become a prevalent part of modern work culture. and it is safe to assume that we currently have more people that are well-versed in the art of remote meetings than ever before.

Notwithstanding, important questions have not been adequately addressed, such as:

- How is the structure and dynamics of remote meetings best described?
- Are they as effective as collocated ones?
- How do they differ?
- How, for that matter, is effectiveness evaluated?

The actualization of these questions has motivated us to study remote meetings, with the aim of describing, analyzing and comparing spoken interaction behaviors and the resulting efficiency in collocated, remote and hybrid meetings.

In this paper, we focus on the corpus construction explaining the recording phase (setting, participants and the tasks) and also the annotation phase providing a detailed description of the used annotation scheme.

1.1 Scope

According to Merriam-Webster, the term "meeting" simply refers to "the act of coming together." Other definitions, such as Google's English Dictionary by Oxford Languages, specify a more deliberate gathering, defined as "an assembly of people for a specific purpose." Here, we are concerned with this latter kind of purposeful meeting - the kind that Goffman considers "the natural unit of social organisation in which focused interaction occurs", where focused interaction is "when people effectively agree to sustain for a time a single focus of cognitive and visual attention, as in a conversation, a board game, or a joint task sustained by a close face-to-face circle of contributors" (Goffman, 1961). We will however interpret "close face-to-face circle" loosely to allow the inclusion of remote and hybrid meetings. Other constraints typically associated with the term meeting include synchronicity among the participants and a limitation in time, denoted by a beginning and an end (Fulk & Collin-Jarvis, 2001).

We further limit the scope to professional meetings, and more specifically those that occur in the segment of the workforce that has been labelled "knowledge workers". The term was first used by Drucker around 1960 (Drucker, 1959, 1961). Although it is not a particularly well-defined concept (Scarborough, 1999), it commonly includes occupations such as doctors, lawyers, scientists and academics. Around the turn of the century, Drucker explicitly included what he labelled "knowledge technologists", exemplified by computer technicians, lab analysts, paralegals,

software designers, into the group (Drucker, 1999).

2. Background and Related Work

2.1 Studies of Meetings

The study of meetings has received attention in several disciplines. A key body of work is what we may term the social psychology of small groups, a field that reached a peak in the 70s. (Davis et al., 1976) presents a comprehensive overview that is particularly relevant. The focus here is not as much on meetings as it is on the dynamics of work in small groups in general, and although the two are clearly associated, they are not the same. Goffman points to several reasons to hold the two concepts apart, with the strongest being that the crucial meeting attribute of “maintenance of continuous engrossment in the official focus of activity” is “not a property of social groups in general” (Goffman, 1961).

In literature more directly focused on meetings there are several directions that are worth specific mention here. The study of the effects of meeting facilitators blossomed in the 1990s, in part because of the increase of team-based organizations, but also because “the advent of group support technologies” (e.g., audio-video conferencing) created “a special demand for facilitation” (Niederman & Volkema, 1999; here you will also find a brief overview of the field). Directly related to the same technology development are studies of the effects of distance (e.g. in audio-video conferencing). (Fulk & Collin-Jarvis, 2001) provides a comprehensive overview of 20th century work in this field.

2.2 Meeting Types

McGrath (1984) takes off from Hackman’s three classes (e.g. Hackman & Morris, 1975; Morris, 1966), where “production” and “problem-solving” becomes generated (ideas and plans, respectively) and “discussion” becomes choose or negotiate depending on the situation. He then adds “execute” as a fourth alternative. Each of these four basic “quadrants” is then divided into two using features from several other classifications. This results in eight task types: planning tasks, creativity tasks, intellectual tasks, decision making tasks, cognitive conflict tasks, mixed-motive tasks, competitive tasks, and psycho-motor tasks. This classification - the circumplex model - has been quite influential. In this terminology, our main interest is in the “choose” quadrant, and more specifically in “decision making tasks”.

2.3 Mediated Meetings

A great deal of theoretical work on mediated meetings took place quite some time ago. There is relevant work in the group decision support systems (GDSS) field, although it targets groups

rather than meetings. DeSanctis & Gallupe, (1987) proposed a division of electronic support systems (for group decisions) into three levels. Level 1 contains “technical features aimed at removing common communication barriers”. This is the most relevant level for the present work, as it contains audio/video conferencing.

Fulk & Collin-Jarvis (2001) makes a three-way distinction between group support systems (GSS, which do not seem to differentiate from GDSS, and which refers to all three levels of DeSanctis & Gallupe). Here, their notion of “teleconferencing”, which refers to “meetings held through audio-conferencing and video-conferencing systems” is the main area of interest from our perspective. Review papers on teleconferencing started appearing as early as the 1970s (Williams, 1977). We note, however, that even though audio and video conferencing technology has improved by leaps and bounds since its infancy, acceptance of distant meetings may not have increased at the same rate (Blenke et al., 2017), at least not before the pandemic. We also note that the main issue may not be acceptance but rather that video conferencing and face-to-face meetings simply work differently (Denstadli et al., 2012) and that attitudes vary with the type of video conferencing system used (Julsrud et al., 2012).

Face-to-face interaction is another research field that has taken a keen interest in video conferencing, targeting its presumed inability to faithfully transfer communicative cues and the resulting deterioration on quality of interaction. Various complex technical video solutions have been proposed from near the dawn of video conferencing until the present (e.g. Adeboye, 2020; Nguyen & Canny, 2007; Okada et al., 1994; Sellen et al., 1992), as well as considerably more complex solutions involving avatars in order to achieve telepresence over low bandwidth (Al Moubayed et al., 2012; Beskow et al., 2009)

Directly related to the same technology development are studies of the effects of distance (e.g. in audio-video conferencing). Fulk & Collin-Jarvis (2001) provides a comprehensive overview of 20th century work in this field.

For more direct comparisons of remote and collocated communication, digital interaction has been shown to reduce perceived social presence between communicators, potentially hindering relationship-building among collaborators, and leading to a stronger focus on self (their personal goals) and less on their interaction partners (Scholl et al., 2020). Collocated interactions involve richer visual, auditory, tactile, and contextual information, helping people pick up important social cues and share intentions and emotions resulting in feelings of social closeness (Newson et al., 2021). More generally, *media richness theory* posits that interactions held through “richer” communication media (i.e. media

that involve more cues) lead to better communication. Neshor Shoshan & Wehrt (2022) showed that meetings held through video conferences cause more exhaustion, indicating that so-called *Zoom fatigue* may objectively exist. Moreover, participants involved in remote meetings described difficulties in reading social cues of others, while perceiving pressure to provide such cues themselves (ibid).

Concerning efficiency, Denstadli et al. (2012) showed that while remote meetings save time (both in planning and in the duration of the meeting itself) they are not suitable for participants who do not know one another beforehand, and it makes developing contacts difficult. From a more organizational point of view, colocated meetings are preferred because of the desire to develop social relations and social capital and to handle tasks with high ambiguity (ibid). Similarly, Alge et al., (2003) examined the effect of teams' past experiences on their ability to communicate in colocated and remote contexts. Results indicate that teams without knowledge-building experience (no shared past) communicating colocated reported higher openness/trust and shared more unique information than remote teams communicating through a synchronous computer-mediated medium.

3. Method

3.1 Corpus Collection

3.1.1 Participants

The corpus consists entirely of three-party conversations. Altogether thirty individuals (13 females and 17 males) participated in this study. They were mainly recruited through the Accindi digital platform where researchers and study participants can interact. They were compensated by four cinema tickets. Participants were between 23 and 48 years old. They were all fluent English speakers and had no hearing problems. They formed groups of three while participating in meetings (three individuals per session having three types of meetings consecutively, no individual took part in more than one session).

3.1.2 Tasks

Three different ranking tasks were used during the meetings: NASA moon survival (Hall & Watson, 1970; Littlepage et al., 1995), Desert survival (Lafferty & Pond, 1974; Littlepage et al., 1995) and the Camping game survival (Hare, 1952). In all three tasks an imaginary situation is explained during which participants must find a way to survive. There is a list of items (10 to 15 items dependent on the task) which could help them in their survival. Participants were asked to rank these items from one to fifteen according to their importance for their survival. The aim of this type of task is to arrive at a group consensus by

the end of the meeting. The reason for choosing these survival tasks was that they were well studied and vastly used in the literature. During each meeting the groups had to complete one task and the order in which the tasks were used was randomised.

3.1.3 Setting & Equipment

All meetings took place in the Division of Speech, Music and Hearing (THM) at KTH. Meetings were performed and recorded (both audio and video) in three different settings. The colocated meeting took place in the seminar room at TMH, where participants gathered around a table working on their task. Their meeting was recorded using the meeting owl pro (360-degree camera, mic, and a speaker) which was placed at the center of the table and connected to a host computer. In addition, separate Xoom voice recorders were used to capture audio of each individual. For the digital and hybrid meetings, Zoom video conferencing software was used and participants were placed in separate booths while connecting over Zoom. They were asked to use full screen mode while selecting the gallery view and "Hide self" in the gallery options. The meetings were recorded both through zoom and voice recorders. During the hybrid meeting, two participants were sharing the same room while the third participant was connected through Zoom.

3.1.4 Process

In each session, participants in groups of three, joined three consecutive meetings (colocated, remote, and hybrid) while working on one of the survival tasks in each meeting. The order in which the meetings took place was randomized for each group. Before the start of each session recording, participants were provided with instructions and asked to sign a GDPR consent form and fill out a demographic form. Each meeting, regardless of the setting, consisted of three phases: the pre-meeting, the in-meeting and the post-meeting phase. Before the meeting (the pre-meeting phase) participants were asked to work on the given task individually and write down their individual preferred order of items. They were given 5 minutes to complete this. During the meeting (the in-meeting phase) they had 15 minutes to discuss the same task with their group mates and come up with a group consensus. After the meeting (the post-meeting phase) they were again given 5 minutes to review their initial individual ranking and modify it if necessary. Each session was completed within 2 hours.

3.2 Corpus Annotation

3.2.1 Data Processing & Annotation Tool

The recordings have been segmented and annotated on various levels using ELAN 6.3 multimodal annotation tool (Sloetjes &

Wittenburg, 2008). With ELAN a user can add an unlimited number of textual annotations to audio and/or video recordings. Annotations can be created on multiple layers, called tiers. Tiers can be hierarchically interconnected (child and parents tier). An annotation can either be time-aligned to the media or it can refer to other existing annotations. The content of annotations consists of Unicode text and annotation documents are stored in an XML format (EAF).

3.2.2 Annotation Scheme

The conversations are annotated on various levels using separate tiers for each layer. On the first level, conversations are manually decomposed into **TURN UNITS**. These units are defined as stretches of speech produced by one participant who occupies the speaker role, bounded by periods of inactivity (i.e. silence) of that speaker. (Brady, 1968; Bunt et al., 2020; Heldner & Edlund, 2010). An annotation segment on a tier starts with the start of the vocalisation by a participant and ends with its end. The minimum silence from a participant required to end a TurnUnit was 500ms. The TurnUnits related to each participant were annotated on separate tiers (**TurnUnit_A**, **TurnUnit_B** and **TurnUnit_C**). If there were any other vocalisations by anyone that isn't one of the participants, it is annotated on a separate tier (**TurnUnit_Other**).

On the second level of annotation, the **FOCUS** tier tracks entities currently under discussion in the conversation. This can be seen as a linear, incremental, and simplified version of the 'questions under discussion' concept (Ginzburg, 2001; Larsson, 2002). The **FOCUS** tier has three child tiers (**ITEMS**, **RANK** and **SPEAKER**). Parent and child tiers are linked in such a way that some changes made on a parent tier will also affect its child tiers (child tiers are shown with the same color, see Figure1).

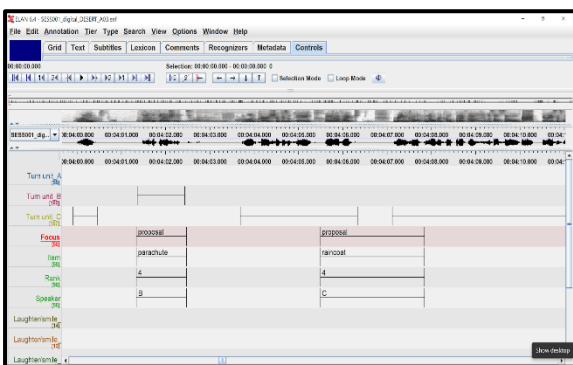


Figure 1: Screenshot of the annotation in ELAN.

Survival tasks only allow for two types of task internal entities to focus on: the **ITEMS** on the list, the **RANK** on which each item belongs. Task externally, we are also interested in which **SPEAKER** is behind an utterance. Anything else

is either not immediately related to the task, or an attribution, argument, etc. that is associated with one or more of these three entities. In our model, each time an item or a rank is mentioned, the entity becomes the focus of its kind. The **ITEMS** tier Shows which item from the list is under discussion (the items are different in different tasks). Focus item changes are defined by a simplistic rule: any mention of an item sets that item in focus. In other words, the mention of a list item sets the item focus to that item. In a similar manner to items, the **RANKS** tier shows which position on the list is considered. In the NASA moon survival and the Desert survival task there are 15 items while in the Boys scout survival task there are only 10 items to rank. And finally, the **SPEAKER** tier shows which speaker made the contribution (is talking).

Furthermore, using controlled vocabulary (CV) in ELAN, we annotate focus-changes as one of **proposal**, **question**, **decision** or **decision-repeat**. When a certain linguistic type with a limited number of annotation values is frequently used it might be a good idea to associate it with a CV. Such a CV consists of a number of predefined values that a user can choose from when editing an annotation, in order to make the task of the annotators less error-prone.

- A **proposal** leaves both Item and Rank set. For example, a participant says: “*I think map (Item) should be in position 2 (Rank)*” or responds “*In the second position*” to the question “*Where should we place the map?*”.
- A **question** sets one of Rank or Item and in effect vacates the other: “*what do you think should go first*” (sets Rank to 1 and Item empty) or “*Where do you think map should be placed?*” (sets Item to map and Rank to empty).
- A **decision** marks the point where the group announces the final consensus on an item and its rank:” *Okay, we put map on fifth position*”.
- A **decision-repeat** marks the instances where a ranking is repeated during the conversation after the decision is made.

Note that in this version of the annotation, we do not annotate grounding and repetitions before the decision at all. The only time we include repetitions (of an already focused Item or Rank, or of a proposed mapping between the two), is when it is a repetition of a decision already made (see decision-repeat above). Other repetitions are simply left unannotated.

LAUGHTER/SMILE are annotated on separate tiers for each speaker when it is audible or visible (Laughter/Smile_A, Laughter/Smile_B and Laughter/Smile_C). The start of the laughter or

smile is marked as the starting point of the annotation segment, and its end is the end of the segment. Laughter is distinguished from smile by a simple token: the former produces an audible sound while the latter does not.

BREATHING (*In_breath* and *out_breath*) were also marked on separate tiers (Breathing_A, Breathing_B and Breathing_C) for each speaker and are annotated only when it's audible. The starting point of the annotated segment is when the breathing begins, and it ends when the breathing ends.

Acoustic **SILENCE** (*SIL*) was defined as a segment in which no participant vocalizes, flanked by segments in which some participant vocalizes. Silences were not annotated explicitly but found by extracting segments with no TurnUnits.

Although annotators were instructed to produce adjacent TurnUnits when no silence could be heard, there were a few mistakes of this sort in the original annotations. In a semi-automatic post-processing step, we removed any within-speaker silences of less than 500 ms (which is the minimum duration of gaps to be annotated according to the annotation scheme). A total of 4% of the automatically extracted silences were removed by this process.

In addition, we removed any between-speaker silences of 50 ms or less. The reasoning here is that these acoustic silences are not perceivable as silences, or gaps, in the terminology of Sacks et al (Sacks et al., 1974). On average, the group decision threshold for perceivable acoustic silence between speakers is considerably longer - 120 ms, but as some listeners perceive gaps robustly at as little as 58 ms of acoustic silence, we opted for a conservative threshold of 50 ms (Heldner, 2011). Whenever a silence was removed, the adjacent TurnUnits were corrected so that they become adjacent to each other, by growing the larger of the TurnUnits. Removing all between-speaker silences below the group decision threshold of ~120 ms, another 4 % of the between-speaker silences were removed.

Finally, a new entity was added: the **Unbroken Speech Sequences (USS)**. This is a continuous sequence during which at least one participant vocalises at each moment, flanked by silence on both sides. Large proportions of overlap, high intensity, and long TurnUnits all contribute to long USSs, whereas large numbers of pauses, short utterances and general inactivity contribute to low USS durations.

3.3 General Statistical Observation

In total, the corpus consists of 6 hours and 20 minutes. The average total active meeting time in a session ranged from 21 to 46 minutes with an average and median of 38 and 39 minutes, respectively. The average single meeting duration

ranged from five to 20 minutes with both average and median at 13 minutes.

In terms of TurnUnits, the corpus contains a total of 8149 TurnUnits. The number of TurnUnits produced by a group (in one session) ranged from 587 to 1120 with an average of 815 and a median of 816 TurnUnits. The number of TurnUnits in a single meeting ranged from 117 to 411 with an average of 272 and a median of 277 per meeting.

The total number of unbroken speech sequences (USS) is 5293, with a range from 328 to 689 in a single session, an average of 530 and a median of 567. That means that a typical USS contained 1.5 TurnUnits (average and median), with a highest session TurnUnits/USS at 2.7 and a lowest of 1.3 (note that the floor is 1, here, as each USS holds at least one TurnUnit). The number of USSs in a single meeting ranged from 79 to 270, with an average of 176 and a median of 180. A USS contains on average 1.6 TurnUnits, while the corresponding median was 1.7, with a lowest observation of 1.1 and a highest of 4.1 TurnUnits/USS.

The median TurnUnit duration in the entire corpus was 1.5 s, and the median silence 0.7 s. Medians within sessions range from 1.1 to 2 s for TurnUnits and from 0.6 to 1 s for silences.

4. Summary & Future Work

We have presented the MEET meeting corpus and its annotation scheme. Ten three-person groups were recorded, each in a single session consisting of three separate meetings, each with a different task and condition, for a total of 30 meetings.

This corpus was constructed with the aim of systematically describing, analysing and modelling interaction patterns during different types of meetings plus evaluating the outcome of these different meeting setups. We wanted to know in which setting the participants were more cooperative and had the highest influence on the group consensus.

Although the current legislation does not permit us to share the corpus recordings, we plan to release the interaction models together with a detailed description of how they were derived. As for future work, we will release tools that facilitate work with and analysis of the kind of interaction model we have created. We will also include more of the annotation, for example filled pause annotation (any spoken sound or word used to fill gaps in speech) in the model. Currently, these are manually segmented for all the meetings in the corpus, but not included since they have not been validated. We also plan to have the corpus transcribed. Currently a section of it is transcribed using whisper ASR. It is however unclear how much of the transcriptions can be shared freely.

5. References

- Adeboye, D. (2020). Exploring the effect of spatial faithfulness on group decision-making (Master Thesis UCAM-CL-TR-952; p. 64). University of Cambridge.
- Al Moubayed, S., Skantze, G., & Beskow, J. (2012). Lip-reading: Furhat audio visual intelligibility of a back projected animated face. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Intelligent Virtual Agents* (pp. 196–203). Springer. https://doi.org/10.1007/978-3-642-33197-8_20
- Alge, B. J., Wiethoff, C., & Klein, H. J. (2003). When does the medium matter? Knowledge-building experiences and opportunities in decision-making teams. *Organizational Behavior and Human Decision Processes*, 91(1), 26–37. [https://doi.org/10.1016/S0749-5978\(02\)00524-1](https://doi.org/10.1016/S0749-5978(02)00524-1)
- Beskow, J., Salvi, G., & Moubayed, S. A. (2009). SynFace—Verbal and non-verbal face animation from audio. *The International Conference on Auditory-Visual Speech Processing AVSP'09*.
- Blenke, L. R., Gosavi, A., & Daughton, W. (2017). Attitudes towards face-to-face meetings in virtual engineering teams: Perceptions from a survey of defence projects. *International Journal of Project Organisation and Management*. <https://www.inderscienceonline.com/doi/abs/10.1504/IJPOM.2017.085284>
- Brady, P. T. (1968). A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal*, 47(1), 73–91. <https://doi.org/10.1002/j.1538-7305.1968.tb00031.x>
- Bunt, H., Petukhova, V., Gilmartin, E., Pelachaud, C., Fang, A., Keizer, S., & Prévot, L. (2020). The ISO standard for dialogue act annotation, second edition. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 549–558. <https://aclanthology.org/2020.lrec-1.69>
- Davis, J. H., Laughlin, P. R., & Komorita, S. S. (1976). The social psychology of small groups: Cooperative and mixed-motive interaction. *Annual Review of Psychology*, 27(1), 501–541. <https://doi.org/10.1146/annurev.ps.27.020176.002441>
- Denstadli, J. M., Julsrud, T. E., & Hjorthol, R. J. (2012). Videoconferencing as a mode of communication: A comparative study of the use of videoconferencing and face-to-face meetings. *Journal of Business and Technical Communication*, 26(1), 65–91. <https://doi.org/10.1177/1050651911421125>
- DeSanctis, G., & Gallupe, R. B. (1987). A foundation for the study of group decision support systems. *Management Science*, 33(5), 589–609. <https://doi.org/10.1287/mnsc.33.5.589>
- Drucker, P. F. (1959). *Landmarks of tomorrow*. Harper.
- Drucker, P. F. (1961). Fifty years of management—A look back and a look forward. *Journal of Engineering for Industry*, 83(3), 366–370. <https://doi.org/10.1115/1.3664530>
- Drucker, P. F. (1999). Knowledge-worker productivity: The biggest challenge. *California Management Review*, 41(2), 79–94. <https://doi.org/10.2307/41165987>
- Fulk, J., & Collin-Jarvis, L. (2001). Wired meetings: Technological mediation of organizational gatherings. In F. Jablin & L. Putnam (Eds.), *The new handbook of organizational communication* (pp. 625–663). SAGE Publications, Inc. <https://doi.org/10.4135/9781412986243>
- Ginzburg, J. (2001). Interrogatives: Questions, facts and dialogue. In *The Handbook of Contemporary Semantic Theory*. Blackwell.
- Goffman, E. (1961). *Encounters: Two studies in the sociology of interaction* (p. 152). Bobbs-Merrill.
- Hackman, J. R., & Morris, C. G. (1975). Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 8, pp. 45–99). Elsevier. [https://doi.org/10.1016/S0065-2601\(08\)60248-8](https://doi.org/10.1016/S0065-2601(08)60248-8)
- Hall, E. J. (Jay), & Watson, W. H. (1970). The effects of a normative intervention on group decision-making performance. *Human Relations*, 23(4), 299–317. <https://doi.org/10.1177/001872677002300404>
- Hare, A. P. (1952). A study of interaction and consensus in different sized groups. *American Sociological Review*, 17(3), 261–267. <https://doi.org/10.2307/2088071>
- Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *The Journal of the Acoustical Society of America*, 130(1), 508–513. <https://doi.org/10.1121/1.3598457>
- Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4), 555–568. <https://doi.org/10.1016/j.wocn.2010.08.002>
- Julsrud, T. E., Hjorthol, R., & Denstadli, J. M. (2012). Business meetings: Do new videoconferencing technologies change communication patterns? *Journal of Transport Geography*, 24, 396–403. <https://doi.org/10.1016/j.jtrangeo.2012.04.009>
- Lafferty, J. C., & Pond, A. W. (1974). The desert survival situation: A group decision making experience for examining and increasing individual and team effectiveness. *Human Synergistics*.
- Larsson, S. (2002). *Issue-based dialogue management [PhD Thesis]*. Department of Linguistics, Göteborg University.

- Littlepage, G. E., Schmidt, G. W., Whisler, E. W., & Frost, A. G. (1995). An input-process-output analysis of influence and performance in problem-solving groups. *Journal of Personality and Social Psychology*, 69(5), 877–889. <https://doi.org/10.1037/0022-3514.69.5.877>
- McGrath, J. E. (1984). *Groups: Interaction and performance*. Prentice-Hall. https://www.jstor.org/stable/2393041?origin=cr_ossref
- Morris, C. G. (1966). Task effects on group interaction. *Journal of Personality and Social Psychology*, 4(5), 545–554. <https://doi.org/10.1037/h0023897>
- Nesher Shoshan, H., & Wehrt, W. (2022). Understanding “Zoom fatigue”: A mixed-method approach. *Applied Psychology*, 71(3), 827–852. <https://doi.org/10.1111/apps.12360>
- Newson, M., Zhao, Y., Zein, M. E., Sulik, J., Dezechache, G., Deroy, O., & Tunçgenç, B. (2021). Digital contact does not promote wellbeing, but face-to-face contact does: A cross-national survey during the COVID-19 pandemic. *New Media & Society*, 14614448211062164. <https://doi.org/10.1177/14614448211062164>
- Nguyen, D. T., & Canny, J. (2007). Multiview: Improving trust in group video conferencing through spatial faithfulness. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1465–1474. <https://doi.org/10.1145/1240624.1240846>
- Niederman, F., & Volkema, R. J. (1999). The effects of facilitator characteristics on meeting preparation, set up, and implementation. *Small Group Research*, 30(3), 330–360. <https://doi.org/10.1177/104649649903000304>
- Okada, K.-I., Maeda, F., Ichikawa, Y., & Matsushita, Y. (1994). Multiparty videoconferencing at virtual social distance: MAJIC design. *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 385–393. <https://doi.org/10.1145/192844.193054>
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735. <https://doi.org/10.2307/412243>
- Scarborough, H. (1999). Knowledge as work: Conflicts in the management of knowledge workers. *Technology Analysis & Strategic Management*, 11(1), 5–16. <https://doi.org/10.1080/095373299107546>
- Scholl, A., Sassenberg, K., Zapf, B., & Pummerer, L. (2020). Out of sight, out of mind: Powerholders feel responsible when anticipating face-to-face, but not digital contact with others. *Computers in Human Behavior*, 112, 106472. <https://doi.org/10.1016/j.chb.2020.106472>
- Sellen, A., Buxton, B., & Arnott, J. (1992). Using spatial cues to improve videoconferencing. *CHI'92*, 651–652. <https://doi.org/10.1145/142750.143070>
- Sloetjes, H., & Wittenburg, P. (2008, May). Annotation by category: ELAN and ISO DCR. *Procs. of LREC'08. The Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf
- Video conferencing market size, share & Covid-19 impact analysis (FBI100293; Fortune Business Insights, p. 160). (2021). *Fortune Business Insights*. <https://www.fortunebusinessinsights.com/industry-reports/video-conferencing-market-100293>
- Williams, E. (1977). Experimental comparisons of face-to-face and mediated communication: A review. *Psychological Bulletin*, 84(5), 963–976. <https://doi.org/10.1037/0033-2909.84.5.963>
- Yan, Z. (2020). Unprecedented pandemic, unprecedented shift, and unprecedented opportunity. *Human Behavior and Emerging Technologies*, 2(2), 110–112. <https://doi.org/10.1002/hbe2.192>

MSNER: A Multilingual Speech Dataset for Named Entity Recognition

Quentin Meeus^{1,2}, Marie-Francine Moens¹, Hugo Van hamme²

20th Joint ACL-ISO Workshop on Interoperable Semantic Annotation

¹ LIIR Lab, Computer Science Dpt., KU Leuven ² PSI, Electrical Engineering Dpt., KU Leuven
Quentin.Meeus@kuleuven.be

Abstract

While extensively explored in text-based tasks, Named Entity Recognition (NER) remains largely neglected in spoken language understanding. Existing resources are limited to a single, English-only dataset. This paper addresses this gap by introducing MSNER, a freely available, multilingual speech corpus annotated with named entities. It provides annotations to the VoxPopuli dataset in four languages (Dutch, French, German, and Spanish). We have also releasing an efficient annotation tool that leverages automatic pre-annotations for faster manual refinement. This results in 590 and 15 hours of silver-annotated speech for training and validation, alongside a 17-hour, manually-annotated evaluation set. We further provide an analysis comparing silver and gold annotations. Finally, we present baseline NER models to stimulate further research on this newly available dataset.

Keywords: Spoken Named Entity Recognition, Spoken Language Understanding, Speech Dataset

1. Introduction

In an increasingly interconnected world where language knows no boundaries, the field of Speech Processing is undergoing a transformative shift towards multilingual applications. One such pivotal area is Spoken Named Entity Recognition (Spoken NER). Named Entity Recognition (NER) is a natural language processing (NLP) task that involves the identification and categorization of named entities within a text, typically into predefined categories such as names of persons, organizations, locations, dates, numerical values, and more. The primary objective of NER is to automatically recognize and extract specific pieces of information from unstructured text, making it easier to analyze and understand the content. NER plays a crucial role in various NLP applications, including information retrieval, question answering, sentiment analysis, and language understanding. In contrast, *Spoken NER* extracts named entities from audio documents, a task that is considerably more challenging. Indeed, aside from the inherent difficulties associated with speech processing, Spoken NER requires not only to identify and classify the entities, but also to transcribe them correctly. Variability in pronunciation, accents, and dialects can make the detection and especially the spelling of named entities very challenging. On the other hand, prosody, intonation and emphasis are cues that may be crucial for NER but are not readily available in written text. Recognizing the pressing need to facilitate cross-lingual research and to provide comprehensive evaluation resources for Spoken NER models, we have undertaken the task of manually annotating the popular speech dataset VoxPopuli's test sets in four

languages: Dutch, French, German, and Spanish. Additionally, we also provide machine-made annotations on the training and validation sets.

In the following sections, we provide a detailed overview of our efforts in the domain of Spoken NER. First, we give an overview of related works and datasets. Then, we introduce the newly annotated dataset and provide information about its size, multilingual coverage, and its potential significance in advancing Spoken NER technology. Additionally, we describe the methodology employed in the dataset's creation, breaking down the annotation process and data preparation. We also introduce the user-friendly annotation interface we've developed for this purpose. Furthermore, we present the results of various experiments and benchmarks conducted using this dataset. These experiments demonstrate its utility in evaluating Spoken NER models across the chosen languages, highlighting its role in advancing research and development in this field.

In summary, this article describes our contributions to the field of multilingual Spoken NER, including the dataset's creation, annotation methodology, and its role in advancing research in this domain.

2. Literature Review

In the field of NLP, there is not one unified label set. Both generic and specialized datasets exist with their own label sets defined. Specialized datasets might cover large amounts of topics with specific vocabulary and entities. For example, a NER system for doctors would include medications, dosages, medical reasons, etc. (Uzuner et al., 2010), and biomedical entities include names

of proteins, chemical, disease, or species (Crichton et al., 2017). Other datasets provide more generic entities that cover broader landscapes. One of the most widely used is CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003), although it comes with only four entity types (LOC, ORG, PER and MISC). OntoNotes v5 enriches this set with 14 more classes (Table 2), to include things such as numbers, dates, and laws. Its high quality makes it one of the most widely used NER datasets, although it only covers three languages: English, Arabic and Chinese. Another notable mention is Tedeschi et al. (2021), which adds a few more generic classes to OntoNotes definitions to cover things such as animal names, diseases, food, and plants, and released a dataset derived from Wikipedia where named entities were annotated automatically with an annotation pipeline that effectively combined pretrained language models and knowledge-based approaches. A follow-up dataset was published covering more languages (Tedeschi and Navigli, 2022).

Currently, we know of only one Spoken NER dataset that is openly distributed as SLUE (Shon et al., 2021). This is an annotated subset of the larger VoxPopuli dataset (Wang et al., 2021), which comprises audio recordings and corresponding transcripts of sessions held in the European Parliament. The annotated portion of the dataset include approximately 25 hours of speech, divided into three subsets: 3/5 for training, 1/5 for validation, and 1/5 for testing purposes. While this initiative is a significant step forward, SLUE exclusively covers the English language. They used the same entities as OntoNotes (Weischedel et al., 2013) although in practice, they combine some types and remove rare ones to produce a new label set (Table 2, Column 2).

Another task in spoken language understanding is similar to Spoken NER: slot filling. This is the identification of information relevant to specific applications, such as flight booking (Hemphill et al., 1990). Although they share many grounds, there is a major difference: slot filling relates to a specific application, and in this regard, covers a much narrower domain than NER, often consisting of short commands for a computer interface (Lugosch et al., 2019; Saade et al., 2018; Bastianelli et al., 2020; Lugosch et al., 2021; Renkens and Van hamme, 2018) or a booking system (Hemphill et al., 1990). Since the vast majority of entity recognition datasets are text-based, the same goes for the applications. Consequently, NER is often framed as a token classification task, where each word or word piece must be assigned an entity type. Since an entity can cover many tokens, the entity classes are redefined in the BIO format, a widely used tagging scheme in NER tasks (Ramshaw and Marcus,

1995). This format provides a structured way to label and distinguish the boundaries of named entities within the text. Each word or token is tagged with one of three labels: “B” marks the beginning, or first word of an entity, “I” indicates the continuation of the named entity and always follows the “B” tag, and “O” is used for words that are not part of an entity. This marker, together with the entity type, makes the target for the classification task. Other annotation schemes are extensions of this (e.g. IO, IOBES, IOE, etc.). The major drawback of the BIO format is its inability to represent nested entities. The modern approach to NER is to add linear layers to a pretrained language model and fine-tune it on the chosen NER dataset. Sometimes, a conditional random field (CRF) (Lafferty et al., 2001) is added to learn the transition probabilities between the label classes (Ushio and Camacho-Collados, 2021). In Spoken NER, the two main approaches are pipeline and end-to-end models. As the name suggests, pipeline models first use automatic speech recognition to transcribe an audio recording, then use NER to predict the entities. In contrast, end-to-end models do not force the model to make hard decisions by choosing one token over another. Instead, it predicts entities directly from the hidden states. Finally, hybrid models or multitask models predict both the entities and the transcriptions simultaneously (Meeus et al., 2023).

| subset | language | duration | size | entities |
|--------|----------|----------|--------|----------|
| train | DE | 224.5 h | 86,410 | 97,492 |
| | ES | 141.5 h | 47,611 | 66,482 |
| | FR | 186h | 65,952 | 80,255 |
| | NL | 38.5 h | 16,533 | 19,566 |
| dev | DE | 4h | 1,610 | 1,880 |
| | ES | 4h48 | 1,529 | 2,094 |
| | FR | 4h22 | 1,527 | 1,884 |
| | NL | 2h16 | 963 | 1,074 |
| test | DE | 5h | 1,966 | 2,061 |
| | ES | 5h | 1,512 | 2,198 |
| | FR | 4h30 | 1,656 | 2,004 |
| | NL | 2h30 | 1,120 | 1,272 |

Table 1: MSNER Dataset statistics

3. Dataset description

The MSNER dataset is an annotated version of the VoxPopuli dataset (Wang et al., 2021) in four languages – Dutch, French, German, and Spanish. VoxPopuli is a collection of recorded sessions from the European Parliament, segmented to contain one or more sentence by one speaker. For each language in scope, we provide three annotated subsets (Table 1): a training and development set with machine-generated “silver” annotations, and a test set with manual “gold” annotations. The subsets

| OntoNotes5 | SLUE | DE | ES | FR | NL | Examples |
|-------------------|--------|-----|-----|-----|-----|--|
| date | WHEN | 307 | 276 | 243 | 113 | 125 years ago, 15 maart, 1815—1830, 1997 |
| time | | 12 | 21 | 10 | 8 | 24 hours, acht uur, de hele dag, mañana |
| cardinal number | QUANT | 136 | 167 | 123 | 91 | 1, 10, 10 miljoen, 11, 11 billion |
| ordinal number | | 82 | 100 | 79 | 45 | First, Ten derde, dritten |
| quantity | | 6 | 2 | 5 | 1 | one and a half meter, two inches |
| money | | 26 | 16 | 18 | 8 | 200 million EUR, Dertig miljoen euro |
| percent | | 21 | 28 | 13 | 22 | 1 procent, 100%, 15 Prozent |
| geopolitical area | PLACE | 259 | 285 | 283 | 176 | Amsterdam, Australië, Barcelona, Belgium |
| location | | 128 | 139 | 214 | 110 | Afrika, Balkanlanden, Europe |
| group | NORP | 229 | 244 | 285 | 213 | African, American, Christian |
| organization | ORG | 621 | 638 | 527 | 362 | Amnesty International, Charlie Hebdo |
| law | LAW | 64 | 108 | 33 | 22 | Paris Accords, US Constitution |
| person | PERSON | 123 | 131 | 100 | 67 | Angela Merkel, Barroso, Beyoncé |
| facility | - | 6 | 2 | 8 | 12 | Guantánamo, White House |
| event | - | 23 | 25 | 21 | 8 | Europees Semester, Rio conferentie |
| work of art | - | 6 | 3 | 4 | 4 | Green Book, Koran |
| product | - | 4 | 1 | 2 | 8 | 2G, 4G, 5G, iPhone |
| language | - | 3 | 12 | 6 | 2 | Latin, Nederlands, Español |

Table 2: Number of annotated entities per entity type in the test sets. Column SLUE correspond to the ‘combined’ entity set proposed by Shon et al. (2021).

of the four languages in scope were annotated according to OntoNotes’ 18 classes. The test sets were manually annotated by the authors following the methodology outlined in Section 4. Each example in the annotated dataset contains the VoxPopuli ID to identify the relevant audio recording in the original dataset, the transcribed sentence and the annotated named entities, that is, the list of entities, each composed of a text and a label component (Figure 1). For the silver label datasets, we also provide a probability score of each predicted entity. We discuss in Section 6 how this number is related to the uncertainty of the model.

We use the 18-classes OntoNotes label set (Weischedel et al., 2013). However, following the example from Shon et al. (2021), we provide annotations by using an alternative label set that combines entity types like places or numbers and discard the rarest classes like languages, events, and work of art (Table 2 Column 2).


| | |
|----------|---|
| ID | 20090423-0900-PLENARY-26-fr_20 090423-21:55:26_4 |
| Audio |  |
| Text | 200 milliards d’euros qu’il faut rapprocher aussi du niveau des déficits des pays européens. |
| Entities | (MONEY, 200 milliards d’euros) (NORP, européens) |

Figure 1: Annotated example

4. Methodology

We provide two kinds of label quality: machine-generated “silver” labels and human-annotated “gold” labels. For obvious reasons, the silver labels are much cheaper and easier to produce. Therefore, we only provide human-made annotations for the test sets, and the training and validation sets annotations are entirely machine-generated. The methodology follows these four broad steps: (1) filtering out recordings without or with misaligned transcripts, (2) generate silver labels for all subsets, (3) manually annotate the test sets and (4) verify the human-made annotations to identify and rectify potential labelling errors. We detail each step in the following paragraphs.

4.1. Filtering

The VoxPopuli dataset contains a few alignment errors between the spoken content and its corresponding transcript. To address this issue, we employed an automatic speech recognition (ASR) system, initially transcribing the spoken utterances and subsequently calculating the word error rate by comparing the ASR-generated sentence to the provided transcript. For this task, we opted for the Whisper large v2 ASR model (Radford et al., 2022), because it showed near state-of-the-art performance across the selected languages. Notably, this model has been meticulously trained on extensive, well-curated data to perform both audio translation and transcription tasks.

For the training and development sets, we filter out examples with a WER larger than 20%, without verifying that the excluded examples were indeed problematic. This discards about 20% of the Ger-

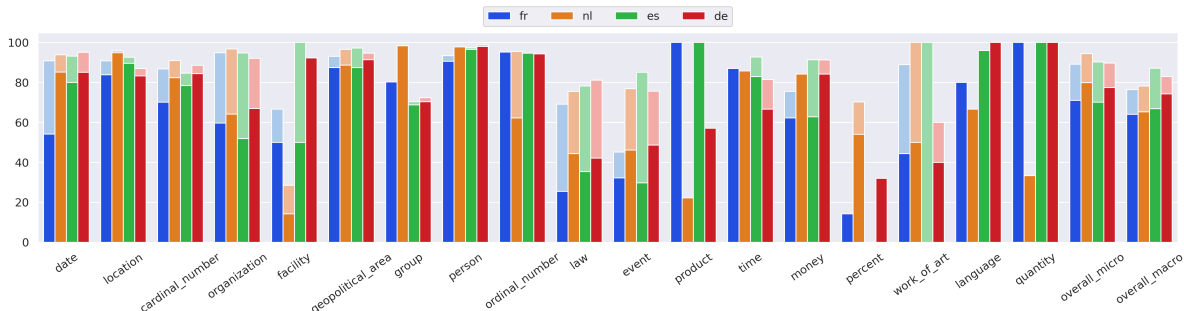


Figure 2: Evaluation of text-based pretrained NER model against our annotations. Bright colors correspond to the F1-score and faded colors correspond to the label-F1 score, a metric that ignores spelling mistakes and segmentation errors.

man and Dutch utterances, 10% of the French examples and 6% of the Spanish utterances.

For the test sets, instances where the word error rate (WER) between the machine-generated transcription and the original transcript exceeded 20%, we conducted a meticulous review process. This involved listening to the audio recording and cross-referencing it with the existing transcript. When feasible, we made necessary corrections to the transcript. However, in cases where multiple speakers were heard in the recording or no speech is present, we removed the problematic utterance from the dataset.

4.2. Pseudo-annotations

We employed an established text-based Named Entity Recognition (NER) model to predict entities within the gold transcript. We chose to use the XLM-RoBERTa large pretrained model (Conneau et al., 2019), fine-tuned specifically on the OntoNotes v5 dataset (Weischedel et al., 2013). This model is readily accessible through the HuggingFace repository¹.

While it’s important to note that this particular model’s fine-tuning was conducted solely on English data, its robustness and efficacy across multiple languages were remarkable. In our evaluation, we observed impressive performance, with most sentences annotated correctly.

4.3. Annotation Tool

For each of the 6,254 pre-annotated sentences in the test sets, we corrected the annotations predicted by the model. For this purpose, we have developed a command line tool to quickly add, edit, merge or remove annotations in a sentence. This utility displays the pre-annotated sentence with a summary of the annotations below. Annotations appear as colored XML tags both in the text and in the summary. An annotated English translation can be displayed. The annotator then has access

¹<https://huggingface.co/asahi417/tner-xlm-roberta-base-ontonotes5>

to both the original sentence and the translation to make sure that the annotations are as accurate as possible. When presented with a sentence, the annotator has the choice to add a new annotation, delete an existing one, merge two annotations together or modify an annotation, either by changing the type or by adding or removing words. Once a sentence has been annotated, it is saved to a file in JSON format. Following this methodology and with the help of this tool, we were able to save a lot of time and effort without sacrificing accuracy. For this reason, we make the tool available online so that others will have the opportunity to contribute to this field of research by easily annotating more data in many more languages.

As mentioned in Section 3, we not only provide annotations according to OntoNotes 18 classes, but also the 7-classes combined set proposed in Shon et al. (2021). However, we chose to completely re-annotate the examples where entities are removed, instead of simply removing all the annotations of the same type from the dataset. To illustrate this, consider the following example:

```
<event> 15th conference on
speech of Toronto </event>
```

According to the combined set conversion rules (Table 2), all the entities of type `<event>` are to be discarded. Doing that would lead to two unannotated entities, ‘15th’ as a number and ‘Toronto’ as a place. Instead, we re-annotate the examples containing removed entities to make sure that we are not penalizing the models for correct assumptions.

4.4. Verification

Finally, we verify the integrity of the test annotations by deriving a number of heuristics and rules that the annotations must abide. This involved grouping the annotations by category and verify each list one by one, comparing them to one another, searching in the text for frequent annotated terms to identify missing annotations, etc. In this last step, we also fix some remaining transcription issues. For example, we realized that VoxPopuli transcripts omitted



Figure 3: Distribution of predicted probability score per class given the target class for the text-based model’s predictions

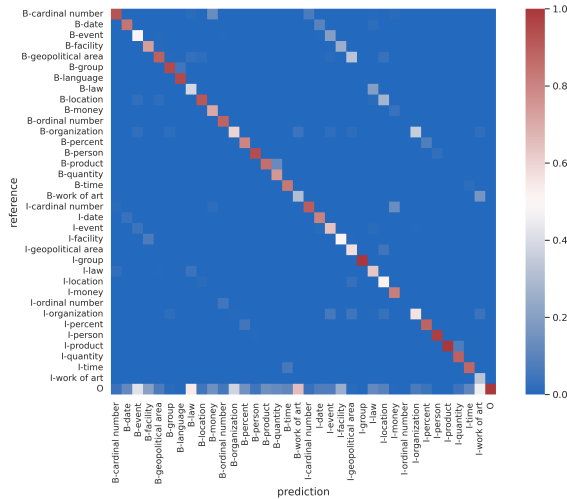


Figure 4: Confusion matrix, normalized to show the probability distribution of the tags predicted with the text-based model.

the symbol “%”, and sometimes the word “thousands” (in all languages). Consequently, for all entities marked as cardinal number, we added the missing tokens when necessary, following the rules specific to the language². Another error often made by the text-based NER model is to predict the article as being part of the entity. As multiple sources advocate against doing so, we abided by the main guidelines (Maekawa, 2018; Benikova et al., 2014).

4.5. Distribution

The annotated datasets are distributed in two formats: As JSON Lines files available on GitHub³, and on the HuggingFace repository (Wolf et al., 2020). There is one file per subset and per language, where each line is an annotated example. The audio files can be obtained by downloading VoxPopuli and matching the audio ID. The dataset version hosted on HuggingFace contains the audio

²In French and in Spanish, the symbol “%” is generally used, but in German and in Dutch, the word is more commonly spelled as Prozent or procent, respectively.

³<https://github.com/qmeeus/MSNER>

recordings and the preprocessed annotations in BIO format, so that a researcher can already use the dataset after only two lines of code.

5. Evaluation Metrics

Following Shon et al. (2021), we recommend evaluating model predictions with the micro-averaged F1-score. The F1-score is the harmonic mean of precision and recall, calculated from an unordered list of named entities predicted for each utterance. Precision is the proportion of correctly predicted entities among all predicted entities, and recall is the proportion of ground truth entities that were correctly detected. An entity is considered to be predicted correctly if both the type and spelling are identical to the ground truth. To allow multiple entities with the same spelling and type in a sentence, we add a unique identifier to each entity/type pair. We recommend using the micro-averaged F1-score because the dataset is unbalanced. The label F1-score only considers the predicted type of the entity for correctness, leaving the transcribed entity out of the computations. This metric ignores spelling mistakes and segmentation errors. We provide an evaluation script⁴ to compute these metrics and generate a breakdown of the prediction results per entity type.

6. Experiments

6.1. Setup

The first analysis compares the annotated test sets to the pseudo-annotations generated by the text-based NER model. Since the silver-label training and validation sets were generated with this model, this analysis is valuable for anyone intending to use these datasets for training. Indeed, it gives insights into the entities that are often confused with one another or remain undetected. It also gives some insights on the reliability of the model’s confidence score in assessing whether a prediction is correct.

⁴<https://raw.githubusercontent.com/qmeeus/MSNER/main/src/evaluate.py>

We also consider two methods to predict named entities from speech, with a pipeline and an end-to-end model. The end-to-end model is a transformer encoder-decoder trained to perform both ASR and NER with a multitask objective (Meeus et al., 2023). This model is initialized from Whisper Large V2 (Radford et al., 2022), with an additional SLU module connected to the layers of the decoder with an adaptor. The end-to-end model was fine-tuned on English SLUE-VoxPopuli (Shon et al., 2021). The pipeline model transcribes the audio files and subsequently annotates the transcriptions. For the ASR model, we use Whisper Large V2 (Radford et al., 2022). For the pipeline model, we provide two options to allow for a better comparison. In Table 3, we use XML-RoBERTa fine-tuned on OntoNotes v5 (Weischedel et al., 2013) and compare it to the predictions generated by the text-based NER model from the gold transcripts. In Table 4, we fine-tuned the same XML-RoBERTa on SLUE-VoxPopuli (Shon et al., 2021), which provides a fair comparison to the end-to-end model. Although both models rely on multilingual pre-trained models, the fine-tuning dataset is entirely in English. Therefore, we evaluate the ability of these models to generalize from one language (English) to other languages (Dutch, French, German, and Spanish). Before computing the F1-scores, we normalize the text by putting it in lower case and removing symbols. It should be noted that the evaluation script does normalize the text further, which could have its importance depending on the model to be evaluated.

All results are presented on the human-annotated test sets proposed in this article.

6.2. Results

Figure 3 shows the distribution of calculated probabilities for predicted ‘B’ and ‘O’ tags conditional to whether they were predicted correctly or not. For each token position k , the probability of the most likely tag i^* is computed as follows:

$$P(y^k = i^*) = \max_i \frac{e^{z_i^k}}{\sum_j e^{z_j^k}}$$

where $z_{1..N}^k$ are the logits predicted by the model for the token at position k . We observe that, on average, annotations for which there was no agreement between the annotator and the NER model were predicted with a lower probability than annotations that were correctly annotated from the start. However, we observe major differences between the class distributions. For the most frequent classes, like ‘O’, ‘organization’ or ‘date’, the probability distributions overlap considerably, and one should be careful if using this score as a proxy for the model’s uncertainty. This is not surprising, as transformers

are known to be overconfident (Ye et al., 2023). For rare quantitative classes like ‘percent’ and ‘quantity’, the model shows confidence when predictions are correct, and uncertain otherwise. This indicates that for those particular classes, the given probability could be relied upon when estimating the model’s uncertainty. The score breakdown by entity and language (Figure 2) indicates that in general, there are no major differences across languages, except for rare classes, where the variability increases significantly.

Figure 4 shows the confusion matrix of the NER model predictions against the manual annotations. Most errors are undetected entities (bottom row in Figure 4) and segmentation errors (I-tags predicted instead of B-tags and inversely, are visible on the lighter diagonals above and below the main diagonal). Some entities remain undetected more often than not, e.g. “work of art” and “event”, which is a sign that predictions are less reliable for these rare classes. Some other types are often confused with one another, like “money” and “cardinal number”. However, all types seem to have at most two confused types. We notice that “geopolitical area” is most often confused with “location” and “law”. In the latter case, this is because many laws are named after cities (e.g. the Paris Agreement, the Warsaw Treaty).

Table 3 compares the text-based NER predictions with the NER predictions obtained from the ASR transcript and generated by the same text-based NER model. The OntoNotes dataset, although in English, provides many well-curated annotations and the NER model trained on this dataset seem to generalize well to the other languages. However, this model was not trained to handle automatic transcripts and we observe a considerable drop in performance when it is asked to process ASR outputs. To make a fair comparison with the end-to-end model, we fine-tune XML-RoBERTa on SLUE-VoxPopuli and report the results in Table 4. The fine-tuning dataset being of much modest size (14.5 hours of training data), the models do not have many examples to learn from. The end-to-end model has a slight advantage because it learns simultaneously the ASR and NER tasks, and it is able to share part of its architecture between both tasks. For example, it seems well able to identify the presence of entities despite a lot of transcription and segmentation errors, as evidenced by the large label F1-score. In contrast, the pipeline suffers much more from the transcription errors because it was pretrained on curated texts and is not expecting noisy ASR transcriptions.

The text-based NER model performs best for Dutch, then German, French and finally Spanish. As the model was trained on English annotations, this ranking is not a surprise, although the ability of the

model to transfer to other languages is impressive. However, for the speech processing models, the same conclusion cannot be drawn. The entity F1-score seem to be correlated with the word error rate, which is influenced by the availability of the different languages in the pretraining set. In other words, for speech models, this is the model’s ability to transcribe foreign languages that will drive the quality of the predictions, rather than how similar the evaluation and the pretraining language are. The label-F1 indicates how accurate a model is at detecting the presence of entity types, disregarding of its ability to transcribe it correctly. Looking at those numbers, we observe again the same behavior as with the text-based entity predictions, namely that entities are more likely to be accurately detected when the evaluation language is more similar to the finetuning language.

| Model | Metric | DE | ES | FR | NL |
|-------|--------------|------|------|------|------|
| Gold | F1 (↑) | 77.4 | 70.1 | 71.1 | 79.9 |
| | Label-F1 (↑) | 89.7 | 90.3 | 89.1 | 94.4 |
| | F1 (↑) | 52.4 | 50.6 | 44.7 | 52.7 |
| ASR | Label-F1 (↑) | 66.2 | 63.6 | 59.4 | 66.1 |
| | WER (↓) | 12.0 | 8.6 | 11.1 | 13.1 |

Table 3: Performance of text-based NER model trained on OntoNotes. Gold corresponds to the model’s predictions from the gold transcripts and ASR corresponds to the model’s predictions on the ASR transcripts.

| Model | Metric | DE | ES | FR | NL |
|----------|--------------|------|------|------|------|
| Pipeline | F1 (↑) | 30.8 | 36.3 | 37.2 | 36.3 |
| | Label-F1 (↑) | 42.7 | 51.6 | 49.5 | 45.9 |
| | WER (↓) | 12.0 | 8.6 | 11.1 | 13.1 |
| End2End | F1 (↑) | 38.3 | 41.3 | 39.6 | 31.2 |
| | Label-F1 (↑) | 76.8 | 77.1 | 78.3 | 78.4 |
| | WER (↓) | 13.3 | 10.5 | 14.5 | 18.2 |

Table 4: Provided baselines on the annotated test sets for a pipeline ASR/NER model and an end-to-end multitask model. Both models were fine-tuned on SLUE-VoxPopuli (Shon et al., 2021)

7. Conclusion

In this manuscript, we have presented MSNER, a new dataset for evaluating multilingual Spoken NER systems. Although NER is a popular topic in NLP, this task has remained mostly unexplored in speech processing and spoken language understanding. To address this issue, we have used a pretrained model to annotate the VoxPopuli training and validation subsets in Dutch, French, German, and Spanish. Additionally, to provide researcher with a gold standard dataset for evaluating their

Spoken NER models, the authors have manually annotated the test sets for these subsets. By analyzing the predictions of a text-based NER model, and comparing them with our annotations, we were able to identify points of attentions for researchers who intend to train a model on silver annotations. For example, in some cases, the model confidence on the predictions can serve as a basis to estimate the correctness of the prediction, but this must be done carefully, since we have seen that transformers can be overconfident. Counter-intuitively, we have shown that most frequent classes are not always the ones where the model’s uncertainty is most reliable. We also looked at the classes that were often confused with one another, which gave us some ideas about which errors might be present in the training and validation sets.

We also provide baselines on the newly annotated evaluation subsets. We selected a pipeline and an end-to-end SLU model, both fine-tuned on English SLUE VoxPopuli (Shon et al., 2021), and we evaluate them on the manually annotated test sets. We saw that in a low resource scenario, the end-to-end model seems to benefit from learning simultaneously to transcribe and to annotate, which allows a better generalization across languages than the pipeline model fine-tuned on the same dataset. Finally, we found that the performance of text-based models on unseen languages is correlated with the similarity of the evaluation language with English. However, for speech models, this is the multilingual transcription accuracy that is the main driver for NER performance. Interestingly, we have seen that the end-to-end model was able to identify the presence of entities much better than the pipeline model, despite a similar overall performance, which illustrate the advantage of sharing parameters across tasks.

8. Bibliographical References

- Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. [SLURP: A spoken language understanding resource package](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. [NoSta-D named entity annotation for German: Guidelines and dataset](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.
- Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. [A neural network multi-task learning approach to biomedical named entity recognition](#). *BMC Bioinformatics*, 18.
- Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania*.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML*. Morgan Kaufmann Publishers Inc.
- Loren Lugosch, Piyush Papreja, Mirco Ravanelli, Abdelwahab Heba, and Titouan Parcollet. 2021. Timers and such: A practical benchmark for spoken language understanding with numbers. *CoRR*, abs/2104.01604.
- Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech Model Pre-Training for End-to-End Spoken Language Understanding. In *Interspeech*.
- Emi Maekawa. 2018. [Annotation guidelines for named entities](#). online.
- Quentin Meeus, Marie-Francine Moens, and Hugo Van Hamme. 2023. [Whisper-slu: Extending a pretrained speech-to-text transformer for low resource spoken language understanding](#). In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *CoRR*.
- Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.
- Vincent Renkens and Hugo Van hamme. 2018. [Capsule networks for low resource spoken language understanding](#). In *Proc. Interspeech*. International Speech Communication Association.
- Alaa Saade, Alice Coucke, Alexandre Caulier, Joseph Dureau, Adrien Ball, Théodore Bluche, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, and Mael Primet. 2018. Spoken language understanding on the edge. *CoRR*.
- Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J. Han. 2021. SLUE: new benchmark tasks for spoken language understanding evaluation on natural speech. *CoRR*.
- Simone Tedeschi, Simone Conia, Francesco Cecconi, and Roberto Navigli. 2021. [Named entity recognition for entity linking: What works and what's next](#).
- Simone Tedeschi and Roberto Navigli. 2022. [Multi-NERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition \(and disambiguation\)](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.
- Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.
- Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010. [Community annotation experiment](#)

for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5).

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP. ACL*.

Wenqian Ye, Yunsheng Ma, Xu Cao, and Kun Tang. 2023. Mitigating transformer overconfidence via Lipschitz regularization. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2422–2432. PMLR.

9. Language Resource References

Ralph Weischedel and Martha Palmer and Mitchell Marcus and Eduard Hovy and Sameer Pradhan and Lance Ramshaw and Nianwen Xue and Ann Taylor and Jeff Kaufman and Michelle Franchini and Mohammed El-Bachouti and Robert Belvin and Ann Houston. 2013. *OntoNotes Release 5.0*. Linguistic Data Consortium LDC2013T19, ISLRN 151-738-649-048-2.

Attitudes in Diplomatic Speeches: Introducing the CoDipA UNSC 1.0

Mariia Anisimova, Šárka Zikánová

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
118 00 Praha, Czech Republic
{anisimova, zikanova}@ufal.mff.cuni.cz

Abstract

This paper presents CoDipA UNSC 1.0, a Corpus of Diplomatic Attitudes of the United Nations Security Council annotated with the attitude-part of Appraisal theory. The speeches were manually selected according to topic-related and temporal criteria. The texts were then annotated according to the predefined annotation scenario. The distinguishing features of the diplomatic texts require a modified approach to attitude evaluation, which was implemented and presented in the current work. The corpus analysis has proven diplomatic speeches to be consistently evaluative, offered an overview of the most prominent means of expressing subjectivity in the corpus, and provided the results of the inter-annotator agreement evaluation.

Keywords: Appraisal theory, diplomatic discourse, corpus linguistics, CoDipA UNSC 1.0

1. Introduction

This paper is aimed at describing the CoDipA UNSC 1.0, a corpus of the thematically and temporally selected diplomatic speeches of the United Nations Security Council (Schoenfeld et al., 2019), annotated with the adaptation of the attitude-part of Appraisal theory (Martin and White, 2005). It describes the annotation scenario together with the annotated data, and offers an overview of the corpus statistics, evaluation of the double annotations, and the future of the project.

The need for such a corpus derives from the specific features of multilateral diplomatic communication, which influence the development of a distinctive type of subjectivity expression, that is rarely addressed.¹

Diplomatic speeches form a distinctive group of texts that are different from other types of discourse in many aspects. These texts are highly *formalized* and *structured*, typically preserving the main outline components in a set order independent of the topic of the meeting or the length of a document.

The syntactic complexity of these texts is mainly dependent on the communicative goal of the speaker, who may either choose shorter and simpler formulations if they wish to be concise and clear or opt for complex syntactic structures and complicated style if their goal is to avoid being specific (Stanko, 2001).

Other prominent characteristics of these texts are the understated tone (Stanko, 2001) and indirect-

ness, which result in implicit formulations, complex syntax, and passivization. These pragmatic features prove to be very important to how diplomats express opinions, which are most frequently not of their own but of the political body they represent (Swain, 2017). It is also because of them, that the diplomatic attitudes require their own approach in the process of annotation.

The format of multilateral communication set in the Security Council does not allow for a direct dialogue between the speakers, causing the argumentation to be rather one-sided and monologic (Swain, 2017).

In our previous publications (Anisimova and Zikánová, 2022; Anisimova, 2021) we have discussed the notion of attitude in diplomatic discourse and described our view on the most suitable annotation schemes for its evaluation, explained the annotation process and environment, as well as the criteria for selecting the data for our corpus of diplomatic speeches. We have then provided the outcomes of the first annotation experiment, which was then utilized for redefining the annotation scenario based on problematic and unclear annotation cases. The described work has led to the creation of the language resource, presented in this paper.

The structure of the paper includes the two main sections, namely:

- Approach, which offers an overview of Appraisal theory, our selected approach to it, the description of the annotation process, and the basic principles of the annotation scenario;
- and Corpus analysis, which provides infor-

¹The corpus and the guidelines are ready for publication after the anonymity period.

mation about the corpus statistics, and inter-annotator reliability.

1.1. Related work

Due to its extensive informativity, Appraisal theory (as described in subsection 2.1) has long been applied to various types of discourses. The detailed description of various aspects of emotionality and opinion makes it useful for both qualitative and quantitative analysis. Appraisal theory is applied in various areas of linguistic research, for instance for analyzing argumentative essays (Lam and Crosthwaite, 2018), literary studies (Busetto and Delmonte, 2019), translation studies (Tajvidi and Arjani, 2017), political (Zhang and Pei, 2018) and diplomatic (Lian, 2018) text analysis, as well as movie, book, and consumer product reviews (Kolhatkar et al., 2020). The extensiveness of the list of possible areas of application corresponds with the versatility of the approach.

Particular practical aspects of annotating appraisal-bearing expressions were described by Read et al. (2007) and Fuoli (2018). In their work, Read et al. (2007) have offered a view on methodology for annotating appraisal, and an overview of the use of this methodology to annotate the corpus of book reviews. An inter-annotator agreement study and the considerations of instances of systematic disagreement are particularly useful for developing an appraisal-related annotation framework.

Another work related to the practical aspects of annotating appraisal was developed by Fuoli (2018). This study offers a step-wise method for the manual annotation of appraisal and covers some of the problematic aspects of this type of annotation, such as challenges in identifying appraisal, challenges in classifying appraisal, and questions of reliability, replicability, and transparency of the annotation process. As for practical applications, one of the bigger available resources is the Simon Fraser University Review Corpus (Kolhatkar et al., 2020) that offers 150 movies, books, and hotel reviews annotated with subjectivity types.

2. Approach

Our approach is based on the attitude part of Appraisal theory (Martin and White, 2005). During the first stage of the corpus creation, we have carried out the first trial annotations and designed the annotation scenario in accordance to the text type and annotation task. After that, the scenario was edited according to the annotators' comments to unify the possible inconsistencies in the approaches to the annotation process. This section provides a description of Appraisal theory, annotation scenario,

and annotation process, as well as the data selection process.

2.1. Appraisal theory

Appraisal theory is an approach to analyzing expressions of subjectivity in a written text (Martin and White, 2005). The theory is located within a framework of Systemic Functional Linguistics (Halliday, 2004), and aims at providing a piece of extensive information about the various types of meanings conveyed by a subjective expression. The three main subsystems of Appraisal theory are:

- **attitude**, referring to feelings as they are construed in texts by distinguishing between emotion, ethics, and aesthetics; the values by which speakers pass judgements and associate emotional/affectual responses with participants and processes;
- **engagement**, providing resources for positioning the speaker's/author's voice with respect to the various propositions and proposals conveyed by a text;
- and **graduation**, describing the resources that allow for graduating the interpersonal impact of an expression (White, 2020).

The framework could be summarized as a comparably extensive tree of choices providing information on various aspects of subjectivity (Taboada, 2017).

2.1.1. Attitude

For the annotation of our corpus, we have selected the attitude part of Appraisal theory. The subsystem of attitude according to Appraisal theory (Martin and White, 2005) provides a framework for the analysis of evaluative expressions by categorizing them into three main attitude types, being an *affect* (an emotional reaction), a *judgement* (a reaction of ethical evaluation), or an *appreciation* (an evaluation of aesthetics), as well as attitude polarity, attitude force, and explicitness. Each category is then subdivided into its own tree of choices making the system a complex and informative structure.

The authors offer a variety of subcategories within each of the types of attitude, which allows for detailed expression of subjectivity. In our approach, we decided to focus on the three main subcategories, namely **affect**, **judgement**, and **appreciation** and their types, as well as categories of **sentiment polarity**, and **explicitness**. Our approach to the attitude framework is presented in Table 1.

In our experience, the range of parts of speech that the attitudes could be expressed with include

mainly adjectives (*proper*), verbs (*violate*), and adverbs (*interestingly*), but also other parts of speech, while the annotated sequence may range from one token to a whole sentence – especially in case an attitude was expressed in an implicit way. However, as per [Martin and White \(2005\)](#) the borders of an attitude may be spread across a discourse unit, irrespective of grammatical boundaries.

| Resource | Type |
|---|--|
| Affect <i>expression of one's feelings</i> | happiness security satisfaction inclination |
| Judgement <i>attitude towards behaviour</i> | normality capacity tenacity veracity propriety |
| Appreciation <i>evaluation of semiotic and natural phenomena</i> | impact quality balance complexity valuation |
| Polarity | positive negative |
| Explicitness | inscribed invoked |

Table 1: Overview of the selected aspects of the attitude system based on [Martin and White \(2005\)](#)

2.2. Approach to data selection

The corpus of annotated speeches consists of 100 texts that were manually selected from the UN Security Council Debates dataset ([Schoenfeld et al., 2019](#)). The language of the data is English, and the speeches were either originally presented in English, or included in a form of the official UN translations. The information about the original language of the speech, as well as the speaker's affiliation and sex, the topic of the session, and its year are stored in the metadata of each text.

The text selection was based on certain criteria, to ensure the data represent diplomatic discourse of the given time period in a balanced way.

The first criterion for the data selection was the *topic* of the meeting. We have decided to focus on international military conflicts at the turn of the century, and among those that are present in the dataset the following topics were selected given their representation within the period of time, covered in the dataset:

- the Palestinian topic, comprising the Israeli–Palestinian conflict;

- the Yugoslavian topic, comprising the meetings dedicated to the Yugoslav wars;
- the Ukrainian topic, comprising the meetings dedicated to discussing the Russo-Ukrainian war;
- the Georgian topic, comprising the War in Abkhazia of 1992-1993, as well as the Russo-Georgian War of 2008;
- and the Iraqi topic, comprising the discussions of the 9/11 terrorist attack (2001) and the subsequent Iraq War, as well as the Gulf War.

Each topic was devoted an equal proportion of space within a corpus, which means that we have selected 20 speeches from the meetings devoted to discussing each of the topics.

The second criterion is connected to the selection of particular meetings that would be representative of the topic. After we have grouped the available speeches and according to the topic, we have selected the meetings that would be included in the corpus. At this stage, our aim was to ensure that the corpus is representative of various stages of each of the included conflicts. Each topic is therefore represented by four sessions of the Security Council, spanning within the given conflict and dataset time frame.

The third criterion in speech selection was the speaker's presumed position towards the topic under discussion. We have differentiated between three types of speakers, namely

- the representatives of the countries that are directly participating in the conflict;
- their allies (if possible among permanent members of the Security Council);
- and a representative of a state, whose international political interests appear to be further from the discussed events (typically among non-permanent UNSC members).

The combination of the three criteria allowed for the creation of a more balanced corpus containing various appraisals of the selected topics, and focused on international armed conflicts of the selected time period.

2.3. Annotation process

1. The first trial annotation was completed by two non-native English speakers with background in linguistics, one of whom is among the authors of the presented paper (annotator A and annotator C). The annotators were instructed to follow the description of attitude subtypes and polarity from [Martin and White \(2005\)](#) and [Oteiza \(2017\)](#). The annotations were conducted following an xml-like

scheme that is described in Table 1, except for the categories of explicitness, which were not yet added to the framework. The achieved dataset consisted of ten speeches with double annotations (around 10000 tokens).

2. This step was followed by calculating the inter-annotator agreement to assess the reliability of the first version of the annotation scenario. The assessment was conducted according to the three levels of depth of the attitude scheme:

- The *complete* agreement, if the annotators agree on the presence of the attitude, attitude-type, subtype, and sentiment polarity on the exact segment of the text;

the F1 for this category is 0.265.

- the *core* category refers to the agreement on levels on annotators agree on the presence of an attitude, and attitude-type;

The F1 for the core agreement is 0.691.

- and the results for the *general* category refer to agreement on the presence of an attitude;

the F1 for the general agreement is 0.713.

Results of this experiment supported the hypothesis that even though subjectivity identification task is complicated there would be quite high agreement between the annotators, whereas the more fine-grained categories may need further development to be understood uniformly.

3. After analyzing the agreement and comparing the annotations, we have proceeded with the creation of the second version of a formal annotation scenario.

4. 80% of data (60490 tokens) was then annotated again by one annotator (annotator A) in the selected environment (see Section 2.3.2) according to the updated annotation guidelines which led to their further improvement. In addition to the above-mentioned improvements, it was decided to add the dichotomous category of explicitness to further enrich the corpus.

5. After the annotation scenario was updated, we have proceeded with the annotation of the whole corpus.

Similar to the very first experiment, the annotations were completed by two annotators with background in linguistics. Both of the annotators are non-native English speakers with a high command of this language.

The main annotator (annotator A) has had the task of annotating the whole dataset (105592 tokens), whereas the annotator B has annotated a smaller subset of texts (ca. 10000 tokens) with the aim of the inter-annotator agreement estimation.

2.3.1. Annotation scenario

The annotation scenario was developed for intrasentential annotation of the attitudes in the diplomatic speeches of the United Nations Security Council. The document provides an extensive step-by-step description, which guides an annotator through the following annotation steps:

- Attitude identification: It first provides two approaches to attitude identification, namely:

1. identifying attitudes by first identifying all of the available subjectivity meanings, which relies on SentiWordNet (Baccianella et al., 2010) for the identification of explicit attitudes;

2. and a context-dependent approach, which requires annotators to first read the whole contextual unit (a sentence) and decide on the presence/absence and the borders of an attitude based on their subjective perception of a text.

This approach allows capturing various implicit expressions of subjectivity as for instance *"That evaluation has been transformed into a brutal reality"*, in example of Judgement, that would be perceived as Affect if not analyzed together with the surrounding context.

- Identification of attitude explicitness

The annotators are asked to distinguish between the explicit and implicit attitudes, as in the following examples of text fragments, annotated with the category of affect: *"We are concerned"* as opposed to *"I would like to use this opportunity to express our serious concern"*.

- Identification of attitude sentiment polarity

At this stage of annotation, the annotators are asked to decide if the attitude conveys positive or negative sentiment, as in the following opposition of positive and negative appreciation excerpted from the corpus: *"the best"* as opposed to *"the most challenging"*.

- Select the appropriate length of the annotated fragment

Depending on the context, it may be necessary to annotate units, which are larger than one token to capture the appraisal-bearing meanings (Read et al., 2007). We advise deciding on the appropriate fragment length based on the attitude explicitness. In this approach, the annotated fragment would either include only the tokens that express the meaning of an attitude in a direct explicit way (the *inscribed* tag), or allow for the inclusion of all

tokens that are required to fully capture the meaning if an attitude is expressed implicitly (the *invoked* tag).

Let us take a look at this distinction by the following examples of the fragments annotated as judgement. The inscribed judgement may take as little as one token to fully capture the meaning (as in *"tireless"*) whereas an invoked judgement requires more context (as in *"The conflicts that have raged over the past few days must be completely stopped"*).

- Select between the three main categories and their subtypes

Annotators were to choose between a variety of categories (first presented in the Table 1). One of the challenges of this type of annotation is the fact that the diplomatic attitudes often differ from the textbook examples (Martin and White, 2005), therefore the annotators were provided with detailed descriptions of attitude, judgement, and appreciation together with their subcategories, as well as the observed doubtful annotation cases.

2.3.2. Annotation tool

During the corpus design stage, we have considered various available annotation tools, which would be compatible with our annotation scenario.

Our initial requirements were:

- support of span annotation, preferably allowing to annotated fragments to overlap as well as span over unannotated tokens;
- support of tree-like annotation schemes;
- convenient import and export of documents;
- support of MacOS or Linux;
- convenient format of the exported documents and annotations, as this would matter at the stage of annotation analysis.

After considering various annotation tools that were available at the time and conducting test annotations, it was decided to proceed with the doccano annotation tool (Nakayama et al., 2018). Doccano is a web-based open-source annotation tool that supports sequence labelling and allows the creation of one's own annotation scenarios. It also provides basic statistics and supports auto-labelling. Another useful feature of this tool is collaborative annotation.

For our project, we have selected the sequence labelling annotation type, together with an additional feature that would allow overlapping entities.

3. Corpus analysis

3.1. Corpus statistics

The corpus consists of one hundred manually annotated speeches, namely of 105592 tokens and 7296 types. The total number of sentences in this corpus is 3296. On average, one text in the corpus consists of 33 sentences, while the average length of a sentence is 32 tokens.

The metadata includes:

- the speaker's **name**;
- their **gender** (title-based distinction);

The ratio of female to male speakers in the corpus is 7 to 93. This study does not focus on gender-specific aspects of the diplomatic discourse, however, our dataset shows, that diplomacy is still a mainly male-dominated area, therefore the number of speeches from women diplomats is much lower.

- the **country** or institution represented;

The full list of all the affiliations alongside the number of their texts is available in Table 2. Most of the speeches belong to permanent UNSC members, such as the United Kingdom, the United States of America, the Russian Federation, France, and China. The dataset includes speeches affiliated with the main countries, connected to the selected topics: Palestine, Georgia, Ukraine, Iraq, and Palestine. It was, however, impossible to include them on the same scale, as none of them are permanent UNSC members.

| Country/Organization | Number of texts |
|---|-----------------|
| Argentina, Brazil, Czech Republic, Secretary-General or Deputy Secretary-General, IAEA Director, Ireland, Italy, Jamaica, Japan, Lebanon, Nigeria, Pakistan, Republic of Korea, Romania, Turkey | 1 |
| Bosnia and Herzegovina, Iraq, Yugoslavia, Croatia, Jordan, Syrian Arab Republic | 2 |
| Israel, Palestine | 3 |
| Georgia | 5 |
| Ukraine | 6 |
| France, China | 9 |
| United Kingdom | 11 |
| United States of America | 12 |
| Russian Federation | 15 |

Table 2: The distribution of countries and organizations within the corpus

- the **topic** of the meeting (as per the UNSC meeting records);
- the **conflict** the meeting was devoted to;
- the **year** of the meeting;

The number of selected speeches in relation to the chronological measures of the UNSC dataset (Schoenfeld et al., 2019) is represented in Figure 1. The corpus aims at covering the time span of the conflicts, although the surge in the number of texts is always connected to real-life events and their discussions.

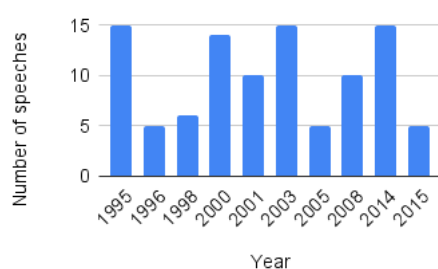


Figure 1: Distribution of the texts over the years

- the meeting identifier;
- the speech identifier.

3.2. Annotation statistics and their interpretation

The following subsection provides information on annotation statistics of the gold data provided by the annotator A.

The total number of attitudes in the CoDipA UNSC 1.0 texts is 1938, with an average of 1.7 attitudes per sentence. The three main categories are represented in the following way:

- Affect: 422 instances
- Judgement: 980 instances
- Appreciation: 536 instances

The average length of an annotated fragment varies from 4.3 tokens for affect to 1.8 tokens for appreciation, and 17.8 tokens for judgement, while the overall average length of a fragment for the corpus is 10.5 tokens. The difference in length corresponds to the preference of either inscribed or invoked modes of expressing subjectivity, with inscribed fragments spanning on average over 2.77 tokens, whereas the invoked fragments - over 16.1.

Let us now take a closer look at the distribution within each of the main categories throughout the corpus.

The distribution of the subcategories of affect is shown in the Figure 2. The two most prevalent categories representing emotional response are *inclination* and *satisfaction*.

Prevalence of the inclination signifies the importance of expressing the diplomats' preferences within the discussed context (if they *incline* and support the events under discussion or other people present during a Council Session), while the category of satisfaction is commonly utilized in the first paragraph of a speech to express the diplomats' emotions towards the other participants of the meeting (with 76.8% of all instances conveying positive sentiment).

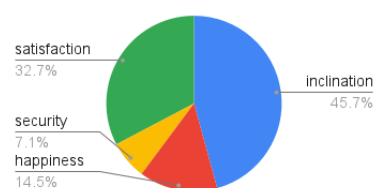


Figure 2: Distribution of the subcategories of affect

The most prevalent subcategory within the judgement subsystem is propriety (Figure 3), which constitutes more than a half of all annotated occurrences. This subcategory is utilized to mark the ethics of the other's (or self) behaviour and belongs to judgement type of social sanction. A curious distribution of the sentiment polarity within the text spans that were identified as propriety (62.7% of tags are positive, and 37.3% are negative) lead us to conclude that the Council members are more interested in advising others on the appropriate course of actions rather than criticize their intentions or behaviour, or praise their and their allies' decisions and actions.

36% of judgements of propriety are formulated by using modalities of ability, permission, obligation, and advice, as in *"the war must be stopped"*, *"our Council should be seized of the matter"*, *"we must demonstrate that we are capable"*.

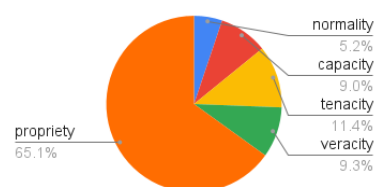


Figure 3: Distribution of the subcategories of judgement

The subcategories of appreciation are represented in a comparatively more diverse way (Figure 4). Here, the three most frequent subcategories are valuation (33.6%), complexity (24.8%), and quality

(20.3%). The entities are being assigned a subjective evaluation based on how valuable they are (value), how well they are put together or are hard to follow (complexity), as well as based on personal preference (quality). In the diplomatic discourse of the UNSC, a part of these expressions is constituted by a set of diplomatic cliches, which repeatedly occurred throughout the corpus (for instance, *"grave consequences"*, *"clear violation"*, *"comprehensive and just solution"*, etc.) and is constituted by a sentiment-bearing adjective.²

The sentiment polarity of the appreciation category is rather positive (68.7% of positive entities and 31.3% of negative entities).

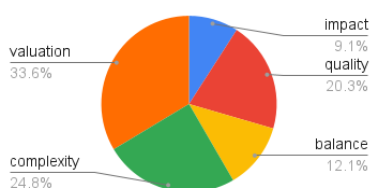


Figure 4: Distribution of the subcategories of appreciation

On a corpus level, the positive evaluations prevail over the negative ones with 64% of all evaluations being positive and 36% being negative. In our opinion, such a sentiment distribution does not completely go in line with the selected topics and could be explained by positive sentiment prevailing in the speeches of the diplomats, who do not represent the countries directly involved in the selected international military conflicts, as well as sufficient amount of subjectivity being directed towards praising themselves and their allies (as in *"the Secretary-General and his Special Envoy have made tremendous efforts"*).

3.3. Inter-annotator agreement

As our corpus was annotated by two annotators, we have selected the Cohen's Kappa as a measure of inter-annotator agreement. The results of the inter-annotator agreement evaluation are presented in the Table 3.

The agreement refers to a sentence-level comparison representing the presence or absence of a designated label in each sentence of a text. The results of the experiments suggest that the inter-annotator agreement is:

- **Fair** for Attitude identification and the subcategory of Affect;

²A diplomatic cliche is an expression that is meant to support the topic of a speech in a standardized way. Such expression could constitute a greeting, an expression of one's condolences, etc.

- **Moderate** for the subcategories of Judgement and Appreciation.³

| | Cohen's kappa |
|--------------|---------------|
| Attitude | 0.41 |
| Affect | 0.44 |
| Judgement | 0.31 |
| Appreciation | 0.32 |

Table 3: Inter-annotator agreement

The agreement level reflects the difficulty of the task as well as the unavoidable subjectivity of attitude evaluation. After a careful manual evaluation of the double annotations,⁴ we have concluded that most cases of disagreement stem from an absence of annotation, which underlines the problematic nature of the attitude identification process as mentioned by Fuoli (2018). However, when annotators do agree on the presence of an attitude in a sentence (with possible small variation in a number of tokens selected) they tend to agree on both on attitude polarity (94.8%), and on the attitude type (79.1%).

3.3.1. Exploring confusion matrices and the main cases of inter-annotator disagreement

Let us take a look at the confusion matrices to summarize the major cases of inter-annotator agreement and disagreement.

As the Figure 5 shows, the annotators mostly agree on the sentences, where they both detect presence of attitudes (93 instances), whereas the biggest disagreement comes from Annotator A not detecting attitudes in sentences, where the Annotator B does (53 instances).

Now, let us illustrate the agreement for three main subcategories of the attitude system with confusion matrices for affect 6, judgement 7, and appreciation 8.

Within the matrices, it is visible that the annotators generally tend to agree on the absence of a tag in a sentence.

The best agreement is observed for the category of affect, while the agreement for the categories of judgement and appreciation is much lower. The reason for the relatively low agreement in general, stems from the subjectivity of the assigned task: it may often appear doubtful which level of semantic meaning should be chosen for the annotation.

Another reason for the lower agreement of the judgement and appreciation is connected to the

³The evaluation of the agreement was derived from the classification described by Koch and Cruz (2004).

⁴The detailed results of the manual analysis of the double annotations will be published separately.

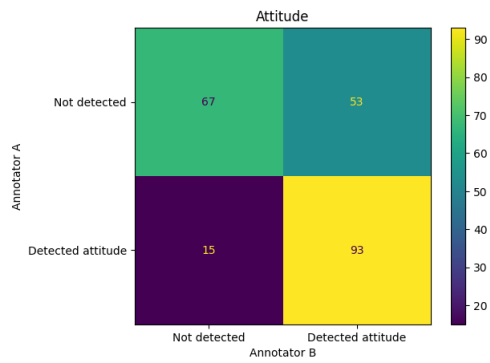


Figure 5: Confusion matrix for the category of attitude

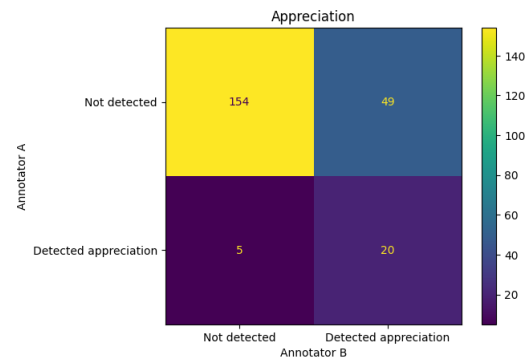


Figure 8: Confusion matrix for the subcategory of appreciation

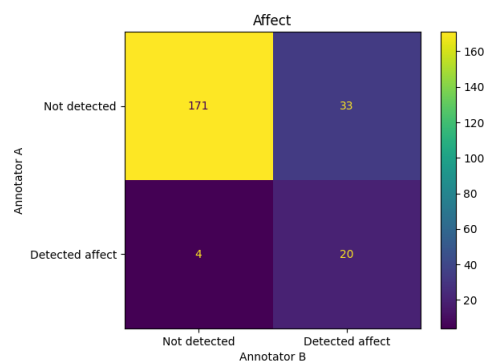


Figure 6: Confusion matrix for the subcategory of affect

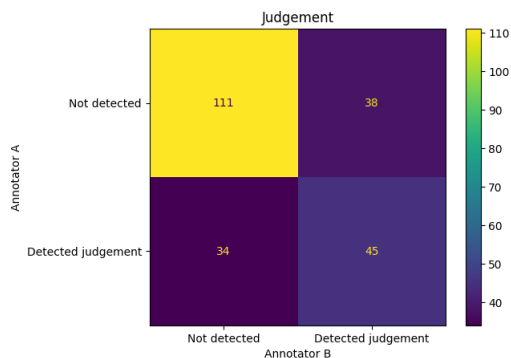


Figure 7: Confusion matrix for the subcategory of judgement

fact that these meanings are often less direct in the diplomatic communication. Affect is still being represented in a way that is quite close to canonical representation of this category (Martin and White, 2005), whereas judgement and appreciation are very often represented in an implicit form, with hidden indirect meanings.

4. Conclusion

This work has introduced the CoDipA UNSC 1.0, a new language resource stemming from Schoenfeld et al. (2019) that provides the data and the framework for analyzing attitudes in diplomatic texts based on Appraisal theory (Martin and White, 2005).

Our corpus offers the annotated dataset that not only proves that the usage of attitudes is consistent throughout the texts, and suggests that the diplomatic texts are highly subjective and evaluative, but also allows for finer-grained attitude analysis based on topical, temporal, and functional criteria.

The most quantitatively significant means of expressing an attitude in the diplomatic speeches of UNSC is judgement, as this category occurs almost two times more often than the other two. The sentiment polarity of the annotations suggests that even though the selected meetings were devoted to discussing the ongoing armed conflicts, diplomats tend on average to keep the positive appearance. This may be explained by various reasons, such as the parties of a conflict being non-prevalent in the corpus, quite significant amount of self-praise, or have other, non-linguistic, explanations.

In our future work, we will focus on further analysis of the obtained language resource from the point of view of possible typical combinations of the attitudinal categories in a text, as well as train a classifier to distinguish between the types of attitudes automatically.

5. Acknowledgements

The research described in this paper has been funded by the Project of the Czech Science Foundation “LuSyD” (No. GX20-16819X), and partially supported by SVV project number 260 698.

6. Ethical considerations and limitations

The constraints of this work lie in its limited scope tied to specific selection criteria, the potential subjectivity inherent in manual annotation, and the likelihood of biases stemming from the selective criteria affecting representational comprehensiveness and introducing variability.

7. Bibliographical References

Mariia Anisimova. 2021. [An introductory overview of evaluating facts and attitudes in diplomatic discourse](#). In *2nd Workshop on Automata, Formal and Natural Languages – WAFNL 2021 Open Session Proceedings*, pages 1–4. Faculty of Mathematics and Physics, Charles University, Prague.

Mariia Anisimova and Šárka Zikánová. 2022. [Attitude in diplomatic speeches: a pilot study](#). In *3rd Workshop on Automata, Formal and Natural Languages – WAFNL 2022 Open Session Proceedings*, pages 151–158.

Nicolò Busetto and Rodolfo Delmonte. 2019. [Annotating shakespeare’s sonnets with appraisal theory to detect irony](#). In *Italian Conference on Computational Linguistics*.

Matteo Fuoli. 2018. [A step-wise method for annotating appraisal](#). *Functions of Language*, 25(2):229–258.

Michael A.K. Halliday. 2004. *An Introduction to Functional Grammar*. Arnold, London, UK.

Sabine Koch and Robyn Cruz. 2004. *Issues of validity and reliability in the use of movement observations and scales*.

Suet Ling Lam and Peter Crosthwaite. 2018. [Appraisal resources in I1 and I2 argumentative essays: A contrastive learner corpus-informed study of evaluative stance](#). *Journal of Corpora and Discourse Studies*.

Yun Lian. 2018. [Analysis of xi’s diplomatic speeches from the perspective of appraisal theory](#). *Journal of Language Teaching and Research*, 9:759–764.

James R. Martin and Peter R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Springer.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang.

2018. [doccano: Text annotation tool for human](#). Software available from <https://github.com/doccano/doccano>.

Teresa Oteiza. 2017. [The appraisal framework and discourse analysis](#). In *The Routledge handbook of systemic functional linguistics*, pages 481–496. Routledge.

Jonathon Read, David Hope, and John Carroll. 2007. [Annotating expressions of appraisal in english](#). In *Proceedings of the Linguistic Annotation Workshop, LAW ’07*, page 93–100, USA. Association for Computational Linguistics.

Nick Stanko. 2001. [Use of language in diplomacy](#). In H. Slavik J. Kurbalija, editor, *Language and Diplomacy*, pages 39–48. DiploProjects, Mediterranean Academy of Diplomatic Studies, University of Malta, Msida MSD 06, Malta.

Elizabeth Swain. 2017. [Interpersonal style\(s\) in diplomatic argumentation online a study of argument schemes and evaluation in press releases of unsc permanent members](#). In G. Garzone C. Ilie, editor, *Argumentation across communities of practice : multi-disciplinary perspectives*, pages 127–148. John Benjamins Publishing Company, Amsterdam, Netherlands.

Maite Taboada. 2017. [System network for appraisal](#).

Gholam-Reza Tajvidi and S. Hossein Arjani. 2017. [Appraisal theory in translation studies: An introduction and review of studies of evaluation in translation](#). *Journal of Research in Applied Linguistics*, 8:3–30.

Peter R. R. White. 2020. [The appraisal website](#).

Siyou Zhang and Zhongwen Pei. 2018. [Analysis of political language based on appraisal theory: The mutual construction of language and powermtaking xi jinping and donald trumprs speeches at world economic forum as examples](#).

8. Language Resource References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. [SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining](#).

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020. [The SFU Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments](#). *Corpus Pragmatics*, 4:155–190.

Mirco Schoenfeld, Steffen Eckhard, Ronny Patz,
Hilde van Meegdenburg, and Antonio Pires.
2019. [The UN Security Council Debates](#).

Automatic Alignment of Discourse Relations of Different Discourse Annotation Frameworks

Yingxue Fu

School of Computer Science, University of St Andrews, Scotland, UK, KY16 9SX
yf30@st-andrews.ac.uk

Abstract

Existing discourse corpora are annotated based on different frameworks, which show significant dissimilarities in definitions of arguments and relations and structural constraints. Despite surface differences, these frameworks share basic understandings of discourse relations. The relationship between these frameworks has been an open research question, especially the correlation between relation inventories utilized in different frameworks. Better understanding of this question is helpful for integrating discourse theories and enabling interoperability of discourse corpora annotated under different frameworks. However, studies that explore correlations between discourse relation inventories are hindered by different criteria of discourse segmentation, and expert knowledge and manual examination are typically needed. Some semi-automatic methods have been proposed, but they rely on corpora annotated in multiple frameworks in parallel. In this paper, we introduce a fully automatic approach to address the challenges. Specifically, we extend the label-anchored contrastive learning method introduced by Zhang et al. (2022b) to learn label embeddings during discourse relation classification. These embeddings are then utilized to map discourse relations from different frameworks. We show experimental results on RST-DT (Carlson et al., 2001) and PDTB 3.0 (Prasad et al., 2018).

Keywords: Discourse annotation, representation and processing, Discourse relations

1. Introduction

Discourse relations are an important means for achieving coherence. Previous studies have shown the benefits of incorporating discourse relations in downstream tasks, such as sentiment analysis (Wang et al., 2012), text summarization (Huang and Kurohashi, 2021) and machine comprehension (Narasimhan and Barzilay, 2015). Automatic discourse relation classification is an indispensable part of discourse parsing, which is performed under some formalisms, the notable examples including the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), based on which the RST Discourse Treebank (RST-DT) is created (Marcu, 1996), and a lexicalized Tree-Adjoining Grammar for discourse (D-LTAG) (Webber, 2004), which forms the theoretical foundation for the currently largest human-annotated discourse corpus—the Penn Discourse Treebank (PDTB) (Prasad et al., 2006, 2018)¹.

As discourse annotation has a high demand on knowledge about discourse, discourse corpora are costly to create. However, these discourse formalisms typically share similar understanding of discourse relations and their role in discourse construction. Thus, an option to enlarge discourse

corpora is to align the existing discourse corpora so that they can be used jointly. This line of work starts as early as Hovy and Maier (1992), but it remains challenging to uncover the relationship between discourse relations used in different annotation frameworks.

Figure 1 shows an example of RST-style annotation. The textual spans in boxes are EDUs and the arrow-headed lines represent asymmetric discourse relations, pointing from satellites to nuclei. The labels *elab(oration)* and *attribution* denote discourse relations. As the two spans connected by the relation *same-unit* are equally salient, the relation is represented by undirected parallel lines. The spans are linked recursively until a full-coverage of the whole text is formed, as shown by the upper-most horizontal line. The vertical bars highlight the nuclei.

As RST-DT and PDTB have an overlapping section of annotated texts, the corresponding PDTB-style annotation on the same text is:

1. *the agreement “an important step forward in the strengthened debt strategy”, that it will “when implemented, provide significant reduction in the level of debt and debt service owed by Costa Rica.”* (implicit, given, Contingency.Cause.Reason)
2. *that it will provide significant reduction in the level of debt and debt service owed by Costa Rica., implemented,* (explicit, when, Temporal.Asynchronous.Succession)
3. *that it will provide significant reduction in the*

¹We focus on RST and PDTB because our method requires a large amount of data and these two frameworks have been applied to the annotation of corpora that overlap in selected texts, thus mitigating the effect of domain shift in the results. Our method does not require corpora built on the same texts.

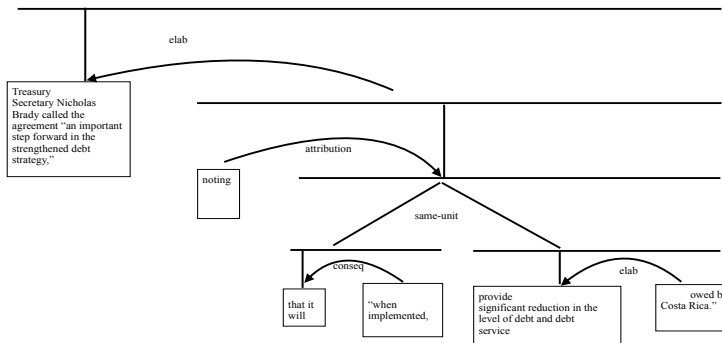


Figure 1: RST-style annotation (wsj_0624 in RST-DT).

level of debt and debt service owed by Costa Rica., **implemented**, (explicit, when, Contingency.Cause.Reason)

where Argument 1 (Arg1) is shown in italics and Argument 2 (Arg2) is in bold. The annotations in parentheses represent *relation type*, which can be implicit, explicit or others, *connective*, which is identified or inferred by annotators to signal the relation, and *sense label*, which is delimited with dots, with the first entry showing the sense label at level 1 (L1 sense), the second entry being the sense label at level 2 (L2 sense) and so on.

The task presents a challenge owing to a multitude of factors. First, different formalisms have distinctive assumptions about higher-level structures and discourse units. PDTB focuses on semantic relations between arguments, and argument identification is performed following the *Minimality Principle*, which means that only those parts that are necessary and minimally required for understanding a relation are annotated (Prasad et al., 2008). In comparison, elementary discourse units (EDUs) in RST are typically clauses. It has been shown repeatedly that segmentation criteria affect the scope of discourse relations and influence the type of relations that can be attached (Demberg et al., 2019; Benamara and Taboada, 2015; Rehbein et al., 2016).

In the first annotation of PDTB, Arg1, i.e., *the agreement "an important step forward in the strengthened debt strategy"*, is taken from the original text "Treasury Secretary Nicholas Brady called the agreement "an important step forward in the strengthened debt strategy"" and the part "Treasury Secretary Nicholas Brady called" is not covered because it does not contribute to the interpretation of the relation here. In contrast, this part is kept in an EDU in RST.

Another major difference between the two frameworks is that RST enforces a tree structure, and all the EDUs and complex discourse units (CDUs) (spans formed by adjacent elementary discourse units and adjacent lower-level spans) should be connected without crossings, while PDTB only focuses on local relations without commitment to any higher-level structure, as exemplified by the three independent annotations shown above. Previous studies (Lee et al., 2006, 2008) suggest that edge crossings and relations with shared arguments are common for PDTB. This distinction adds to the difficulty of exploring correlations of relations between the two frameworks, even if the two corpora are built on the same texts.

In addition, in RST-DT, an inventory of 78 relations is used, which can be grouped into 16 classes. These relations can be divided into *subject matter* relations (informational relations in Moore and Pollack (1992)), which are relations whose intended effect is for readers to recognize them, and *presentational* relations, which are intended to increase some inclination in readers (Mann and Thompson, 1988) (intentional relations in Moore and Pollack (1992)). For each relation, only one sense label can be attached. In contrast, PDTB adopts a three-level sense hierarchy, and more than one sense label can be annotated for a pair of arguments. As shown in the example, annotation 2 and annotation 3 are annotations for the same argument pair, but different sense labels are assigned. In previous studies that explore the alignment of RST and PDTB discourse relations, these cases typically require manual inspection to determine the closest matching PDTB relation to RST (Demberg et al., 2019). Moreover, PDTB does not take intentional relations into account but focuses on semantic and pragmatic relations.

The combination of these factors makes it challenging to investigate the relationship between discourse relations of different annotation frameworks. Even in empirical studies that make use of corpora annotated in multiple frameworks in parallel, expert knowledge and manual examination are still required. To tackle the challenge caused by differences in discourse segmentation, Demberg et al. (2019) employ the strong nuclearity hypothesis (Marcu, 2000)² to facilitate the string matching process of aligning PDTB arguments and RST segments. While this method alleviates the limitation of exact string matching of arguments/EDUs, it relies on a corpus annotated with multiple frameworks in parallel. Furthermore, it is conceivable that the relations left out in their analysis because of violating the principle of strong nuclearity hypothesis are not necessarily irrelevant for the goal of enabling joint usage of RST and PDTB.

In this study, we propose a fully automatic method for this task. We take inspiration from advances in label embedding techniques and an increasing body of research endeavors to harness label information in representation learning, such as supervised contrastive learning (Khosla et al., 2020; Gunel et al., 2021; Suresh and Ong, 2021). Instead of using string matching to identify the closest PDTB arguments and RST EDUs with the aim of discovering potentially analogous relations, we try to learn label embeddings of the relation inventories and compare the label embeddings.

Our contributions can be summarized as follows:

- We propose a label embedding based approach for exploring correlations between relations of different discourse annotation frameworks. The method is fully automatic and eliminates the need of matching arguments of relations.
- We conduct extensive experiments on different ways of encoding labels on RST-DT and PDTB 3.0.
- We develop a metric for evaluating the learnt label embeddings intrinsically and perform experiments to evaluate the method extrinsically.

2. Related Work

Mapping discourse relations Existing research on mapping discourse relations of different frameworks can be categorized into three types (Fu, 2022): a. identifying a set of commonly used relations across various frameworks through analysis of definitions and examples (Hovy and Maier, 1992; Bunt and Prasad, 2016; Benamara and Taboada,

2015); b. introducing a set of fundamental concepts for analyzing relations across different frameworks (Chiarcos, 2014; Sanders et al., 2018); c. mapping discourse relations directly based on corpora annotated in multiple frameworks in parallel (Rehbein et al., 2016). The third approach is closer to our method, and we summarize studies in this direction here. Rehbein et al. (2016) compare coherence relations of PDTB and CCR frameworks on the basis of a spoken corpus annotated in the two frameworks. They find that differences in annotation operationalisation and granularity of relation definition lead to many-to-many mappings. Demberg et al. (2019) show similar findings when mapping relations of RST-DT and PDTB 2.0. To mitigate issues caused by segmentation differences, they use the *strong nuclearity hypothesis* (Marcu, 2000) so that relations that have greater scope than two adjacent EDUs can be covered in their studies. With this method, Costa et al. (2023) maps RST with PDTB 3.0. Scheffler and Stede (2016) propose a method of mapping RST and PDTB relations on a German corpus annotated according to both frameworks. Explicit connectives in PDTB are used as anchors of relations, with some exceptions. It is found that 84.4% of such PDTB explicit connectives can be matched to an RST relation. The results are not surprising, as phrases that begin with a strong discourse marker are specified as EDUs (Carlson and Marcu, 2001), and a relation is likely to be attached. Stede et al. (2016) annotate a corpus with discourse information in RST and SDRT and argumentation information. A set of rules are applied to harmonize the segmentations, and structural transformation into a common dependency graph format is performed. Bourgonje and Zolotareno (2019) try to induce PDTB implicit relations from RST annotation. Segmentation differences present a challenge, and even if the two annotations overlap in segmentation in some cases, different relations are annotated. This observation is consistent with Demberg et al. (2019).

Label embeddings Label embeddings have been proven to be useful in CV (Akata et al., 2016; Palatucci et al., 2009; Zhang et al., 2022a) and NLP tasks (Wang et al., 2018; Zhang et al., 2018; Miyazaki et al., 2019). Conventionally, one-hot encoding is used to represent labels, which suffers from three problems: lack of robustness to noisy labels (Gunel et al., 2021), higher possibility of overfitting (Sun et al., 2017) and failure to capture semantic correlation of labels. Learning meaningful label representations is helpful for mitigating these problems and the semantics of labels can be used as additional information to improve model performance. It is shown that label embeddings are effective in data-imbalanced settings and zero-shot learning (Zhang et al., 2022b).

²A relation that holds between two spans should also hold between the nuclei of the two spans.

Label embeddings can be representations from external sources, such as BERT (Xiong et al., 2021), or can be randomly initialized (Zhang et al., 2022b). Another approach is to learn label embeddings during model training. Akata et al. (2016) propose a method of learning label embeddings from label attributes while optimizing for a classification task. Wang et al. (2018) introduce an attention mechanism that measures the compatibility of embeddings of input and labels. Additional information can be incorporated in learning label embeddings, such as label hierarchy (Chatterjee et al., 2021; Zhang et al., 2022a; Miyazaki et al., 2019) and textual description of labels (Zhang et al., 2023).

3. Method

Problem statement Given a corpus annotated in one discourse annotation framework $D_1 = \{X_m, Y_m\}_{m=1}^M$ and another corpus annotated in a different annotation framework $D_2 = \{X_n, Y_n\}_{n=1}^N$, where X denotes input sequences formed by pairs of arguments, $X_i = A_1^{(1)} \dots A_a^{(1)}, A_1^{(2)} \dots A_b^{(2)}$, and Y represents relation label sets of the two frameworks, $Y_{D_1} = \{y_1, y_2, \dots, y_k\}$ and $Y_{D_2} = \{y_1, y_2, \dots, y_c\}$. The task is to learn a correlation matrix R between Y_{D_1} and Y_{D_2} , which is a $2d$ matrix of shape $k \times c$. Our method is to learn embeddings for members of Y_{D_1} and Y_{D_2} and the widely used cosine similarity can be used as a measure of distance between the embedding vectors. The label embedding learning method is the same for D_1 and D_2 and we use D_1 as an example in the following.

We apply the vanilla version of label-anchored contrastive learning in Zhang et al. (2022b) as the backbone. For an input sequence X_i , we use a pre-trained language model as the input encoder f_{InEnc} . Without losing generality, we choose the popular *bert-base-uncased* model from the Huggingface transformers library (Wolf et al., 2020). For X_i pre-processed as $X_i = [CLS] A_1^{(1)} \dots A_a^{(1)} [SEP] A_1^{(2)} \dots A_b^{(2)} [SEP]$, the representation of the $[CLS]$ token is used as the representation of the input sequence:

$$\mathbf{E}_{X_i} = f_{InEnc}(X_i) \quad (1)$$

where the input sequence representation \mathbf{E}_{X_i} is of shape $(a+b+3) \times dim$, where dim is the dimension of the output from the language model and a and b are the maximum lengths that the arguments are padded to. We empirically find that removing the non-linear transformation to \mathbf{E}_{X_i} in Zhang et al. (2022b) yields better performance for our task.

We explored different options of label encoders, including: adding a BERT model (Devlin et al., 2019) (*LbEncBert*); using a RoBERTa model (Liu et al., 2020), which is trained with

the next sentence prediction objective removed (*LbEncRoberta*); randomly initializing from a uniform distribution (*LbEncRand*); adding text description of the labels (*LbEncDesc*), where the label and the description are processed in the form $[CLS]label[SEP]description[SEP]$, and the representation of $[CLS]$ is used as the label representation; and adding sense hierarchy information, where we use the hierarchical contrastive loss proposed by Zhang et al. (2022a) and apply different penalty strengths to losses at different levels (*LbEncHier*). As we use language models or trainable layers as label encoders, the label embeddings are learnable.

With a label encoder g_{LbEnc} , for k total relations in D_1 , we obtain a table T of shape $k \times lbDim$, where $lbDim$ is the output dimension of the label encoder. Thus, for a label $y_{l=1}^k$, its label embedding vector \mathbf{E}_{y_l} is the $(l-1)^{th}$ row of T .

Instance-centered contrastive loss We apply the method in Zhang et al. (2022b) to compute the instance-centered contrastive loss \mathcal{L}_{ICL} :

$$\mathcal{L}_{ICL} = -\frac{1}{N} \sum_{X_i, Y_i} \log \frac{e^{\Phi(\mathbf{E}_{X_i}, \mathbf{E}_{Y_i})/\tau}}{\sum_{1 \leq l \leq K} e^{\Phi(\mathbf{E}_{X_i}, \mathbf{E}_{Y_l})/\tau}} \quad (2)$$

where N denotes batch size, X_i is an instance in a batch, and Y_i is its label, Φ represents a distance metric between the representations of the input and label embeddings, and cosine similarity is used in the experiment. τ denotes the temperature hyperparameter for scaling, and lower values of τ increase the influence of hard-to-separate examples in the learning process (Zhang et al., 2021). By minimizing this loss, the distance between instance representations and the corresponding class label embeddings is reduced, resulting in label embeddings that are compatible with input representations.

Label-centered contrastive loss The purpose of this loss is to reduce the distance between instances that have the same labels. For a batch with a set of unique classes C , c represents a member, P_c denotes the set of instances in a batch that have the label c and N_c represent the set of negative examples for c . A member in P_c is represented by X_p and a member in N_c is denoted by X_n . The label-centered contrastive loss \mathcal{L}_{LCL} can be computed with:

$$\mathcal{L}_{LCL} = -\frac{1}{C} \sum_{c \in C} \sum_{X_p \in P_c} \log \frac{e^{\Phi(\mathbf{E}_{X_p}, \mathbf{E}_c)/\tau}}{\sum_{X_n \in N_c} e^{\Phi(\mathbf{E}_{X_n}, \mathbf{E}_c)/\tau}} \quad (3)$$

As indicated in Zhang et al. (2022b), \mathcal{L}_{ICL} and \mathcal{L}_{LCL} mitigate the small batch size issue encountered in other types of contrastive learning, which

makes them suitable for scenarios with limited computational resources.

We add the following two supervised losses in the training objective, which we find effective empirically.

Label-embedding-based cross-entropy loss

As shown in Equation 4, a softmax function is applied to the k label embeddings in T , yielding a probability distribution over the k classes:

$$p(y_l) = \frac{e^{\mathbf{E}_{y_l}}}{\sum_{l=1}^K e^{\mathbf{E}_{y_l}}} \quad (4)$$

Let t_{y_l} denote the categorical encoding of the target y_l . The cross-entropy loss of classification based on label embeddings, denoted by \mathcal{L}_{LEC} , can be obtained with Equation 5:

$$\mathcal{L}_{LEC} = - \sum_{l=1}^K t_{y_l} \log p(y_l) \quad (5)$$

The purpose of adding this loss is to make the label embeddings better separated from each other.

Canonical multi-class cross-entropy loss We add the canonical cross-entropy loss for multi-class classification with input representations:

$$\mathcal{L}_{ICE} = - \sum_{i=1}^N \sum_{l=1}^K c_l^i \log p(c_l^i) \quad (6)$$

where N is the batch size, K is the total number of classes, and $p(c_l^i)$ is the probability predicted for a class c . With this loss, the input representations are learnt to be effective for the classification task.

The total loss is the sum of the four losses. During inference, only vector matching between the representation of an input sequence \mathbf{E}_{X_i} and the k learnt embeddings \mathbf{E}_{y_l} is needed, with the cosine similarity as a distance metric, for instance.

$$\hat{y} = \underset{1 \leq l \leq k}{\operatorname{argmax}} (\Phi(\mathbf{E}_{X_i}, \mathbf{E}_{y_l})) \quad (7)$$

Baseline for relation classification We run the *BertForSequenceClassification* model from the Transformers library as the baseline for discourse relation classification, which is trained with cross-entropy loss only, i.e. Equation 6.

Baseline for label embedding learning Label embeddings are generally used for improving performance in classification tasks in previous studies (Wang et al., 2018; Zhang et al., 2018; Xiong et al., 2021; Zhang et al., 2022b). To compare with a method targeted at learning good label embeddings, we implement a baseline method, which is a combination of Equation 4 and 5, but a softmax function is applied over the cosine similarities of an input \mathbf{E}_{X_i} and each label embedding \mathbf{E}_{y_l} in T here, similar to the approach adopted in Zhang et al. (2018) and Wang et al. (2018).

Metric After the model training stage, as the representations of the input sequences have been learnt for the relation classification task, we can leverage the average of the representations of input sequences X that belong to a class y_l as a proxy for the class representation, denoted by \mathbf{H}_{y_l} :

$$\mathbf{H}_{y_l} = \frac{1}{C} \sum_{i=1}^C \mathbf{E}_{X_i} \quad (8)$$

where C represents the number of instances in X .

Due to inevitable data variance, the learnt label embeddings \mathbf{E}_{y_l} for a class y_l may not be the same as \mathbf{H}_{y_l} , but it should have a higher correlation with \mathbf{H}_{y_l} than label embeddings of the other classes. Hence, we compute the correlation matrix M between the k learnt label embeddings \mathbf{E}_{y_j} and the k class representation proxies \mathbf{H}_{y_i} , where $0 \leq j, i \leq k - 1$, with cosine similarity as the metric of correlation:

$$M_{ij} = \Phi(\mathbf{H}_{y_i}, \mathbf{E}_{y_j}) \quad (9)$$

For each class representation proxy, we normalize its correlation scores with the k learnt label embeddings to a range of $[0, 1]$. The average of values at the main diagonal of M is adopted as an overall measure of the quality of the learnt label embeddings:

$$\mathcal{L}EQ = \frac{1}{K} \sum_{i=0}^{K-1} \tilde{M}_{ii} \quad (10)$$

Figure 2 shows the method of intrinsic quality estimation for learnt label embeddings.

| | \mathbf{E}_1 | \mathbf{E}_2 | \mathbf{E}_k |
|----------------|------------------------------------|------------------------------------|------------------------------------|
| \mathbf{H}_1 | $\cos(\mathbf{E}_1, \mathbf{H}_1)$ | $\cos(\mathbf{E}_2, \mathbf{H}_1)$ | $\cos(\mathbf{E}_k, \mathbf{H}_1)$ |
| \mathbf{H}_2 | $\cos(\mathbf{E}_1, \mathbf{H}_2)$ | $\cos(\mathbf{E}_2, \mathbf{H}_2)$ | $\cos(\mathbf{E}_k, \mathbf{H}_2)$ |
| \mathbf{H}_k | $\cos(\mathbf{E}_1, \mathbf{H}_k)$ | $\cos(\mathbf{E}_2, \mathbf{H}_k)$ | $\cos(\mathbf{E}_k, \mathbf{H}_k)$ |

Figure 2: Illustration of the correlation matrix M . $\mathbf{E}_{1 \dots k}$ represents the k learnt label embeddings and $\mathbf{H}_{1 \dots k}$ denotes the k class representation proxies. After normalization, the average of the values at the diagonal (colored) is the overall measure of the quality of the learnt label embeddings.

4. Experiments

4.1. Data Preprocessing

For the purpose of our research, it would be ideal to learn label embeddings for all the relations. However, the label embeddings are trained together

with input representations in a multi-class classification task and data imbalance poses a challenge. Therefore, we focus on 16 relations for RST and PDTB L2 senses with more than 100 instances, following Kim et al. (2020).

The RST trees in RST-DT are binarized based on the procedure in Ji and Eisenstein (2014) and the spans and relations are extracted. The 78 relations are mapped to 16 classes based on the processing step in Braud et al. (2016)³. We take 20% from the training set of RST-DT for validation purpose.

For PDTB, we take sections 2-20 as the training set, sections 0-1 as the development set, and sections 21-22 as the test set, following Ji and Eisenstein (2015).

4.2. Hyperparameters and Training

We run each model three times with different random seeds and report the mean and standard deviation of the results. We use the AdamW optimizer (Loshchilov and Hutter, 2019) and clip L2 norm of gradients to 1.0. The learning rate is set to $1e - 5$. The batch size is set to the maximum that the GPU device can accommodate. The total training epoch is set to 10 and we adopt early stop with patience of 6 on validation loss.

The temperature τ for instance-centered contrastive loss and label-centered contrastive loss is set to 0.1. For the experiment with *LbEncHier* label encoder, the penalty factor is $2^{1/2}$ for L1 loss and 2 for L2 loss.

The learning rate for the baseline *BertForSequenceClassification* model is set to $5e - 5$.

Our implementation is based on the PyTorch framework (Paszke et al., 2019) and a single 12GB RTX3060 GPU is used for all the experiments.

4.3. Results

Since we observe minimal discrepancies in data distributions between the training and test sets, we opt to utilize the test set for generating the class representation proxies necessary for the computation of the metric.

Table 1 shows the experimental results for PDTB and RST. Explicit and implicit relations for PDTB are combined. After the preprocessing step, 16 relations remain for both PDTB and RST.

It can be observed that the performance of label embedding learning on RST is lower than PDTB. Moreover, adding label embeddings generally lowers F1 compared with training with cross-entropy loss only. The decrease in F1 might be related to data sparsity when more learning objectives are

³https://bitbucket.org/chloebt/discourse/src/master/preprocess_rst/code/src/relationSet.py

added but the data amount is the same, which is visible when supplementary information of labels is added, as shown by cases of *LbEncDesc* and *LbEncHier*. This phenomenon is rather pronounced for RST, which has a much smaller data amount. Additionally, although the label encoder *LbEncRand* works best for the classification task, the learnt label embeddings rank the lowest among the different options. Through examination, we find that with this approach, the label embeddings of different classes are not close to the class representation proxies and we conjecture that during training, the label embeddings are mainly used as anchors, as in Zhang et al. (2022b), but the input representations are better learnt, hence the higher classification accuracy and F1 score. Zhang et al. (2022b) did not report other options of label encoders than random initialization and their focus is classification accuracy.

4.4. Data Augmentation for RST

To improve the performance on RST, we use back translation as a means of data augmentation. We translate all the files containing EDUs in the training set (only) into French and translate the French texts back into English, using Google Translate⁴. Data augmentation is not performed for *Elaboration* and *Joint*, which are the two largest classes in RST-DT, to achieve a more balanced data distribution.

Based on the results shown in Table 1, we choose *LbEncRoberta* in the following experiments because of its good performance but results with *LbEncBert* are comparable.

Table 2 shows the results. The F1 scores and label embedding scores are improved to a large margin. As back translation is performed at the EDU level, it is unavoidable that errors are introduced, and given that data augmentation is not performed for the two largest classes, their influence on the results is reduced, hence the lower classification accuracy.

| | Acc. | F1 | Label emb. |
|-------|--------------------|--------------------|--------------------|
| +aug. | 62.75(\pm 0.79) | 50.76(\pm 0.94) | 92.96(\pm 0.90) |
| -aug. | 65.20(\pm 0.07) | 45.39(\pm 0.60) | 76.56(\pm 0.85) |

Table 2: Results for RST with data augmentation (+aug) and without data augmentation (-aug).

Figure 3 shows the T-SNE visualization plots of learnt label embeddings together with the class representation proxies for the test set of RST-DT. The label embeddings learnt with data augmentation are shown in Figure 3a in comparison with Figure 3b, where no data augmentation is performed. It is visible that in Figure 3a, more label

⁴<https://translate.google.com/>

| Data | Label enc. | Acc. | F1 | Label emb. |
|------------|---------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| PDTB total | <i>LbEncBert</i> | 69.45(\pm 0.18) | 57.80(\pm 0.85) | 93.84(\pm 0.37) |
| | <i>LbEncRoberta</i> | 69.34(\pm 0.46) | 58.10(\pm 0.15) | 94.23(\pm 0.74) |
| | <i>LbEncRand</i> | 69.87(\pm 0.80) | 59.00(\pm 0.62) | 89.32(\pm 0.01) |
| | <i>LbEncDesc</i> | 69.16(\pm 0.26) | 57.53(\pm 0.14) | 93.58(\pm 0.42) |
| | <i>LbEncHier</i> | 69.21(\pm 0.45) | 56.70(\pm 0.14) | 93.67(\pm 0.23) |
| | <i>Baseline</i> | 69.42(\pm 0.46) | 58.73(\pm 0.78) | 79.15(\pm 2.06) |
| RST | <i>LbEncBert</i> | 64.62(\pm 0.90) | 44.86(\pm 1.85) | 78.64(\pm 1.02) |
| | <i>LbEncRoberta</i> | 65.20(\pm 0.07) | 45.39(\pm 0.60) | 76.56(\pm 0.85) |
| | <i>LbEncRand</i> | 65.09(\pm 0.70) | 45.53(\pm 4.82) | 69.98(\pm 3.10) |
| | <i>LbEncDesc</i> | 64.62(\pm 0.21) | 43.69(\pm 1.20) | 74.18(\pm 0.91) |
| | <i>LbEncHier</i> | 63.66(\pm 0.50) | 41.30(\pm 0.39) | 74.54(\pm 0.77) |
| | <i>Baseline</i> | 63.55(\pm 0.23) | 48.57(\pm 0.73) | 48.21(\pm 1.27) |

Table 1: With results over three runs collected, the Pearson correlation coefficient between classification accuracy and label embedding scores is 0.5814 and it is 0.8187 between f1 and label embedding scores, both with $p < 0.05$), which shows that the learnt label embeddings are closely related to F1 scores.

embeddings fit into the class representation proxies while in Figure 3b, label embeddings of only six classes are close to the class representation proxies, and the rest form a nebula, which suggests that the label embeddings cannot be distinguished clearly from each other. In Figure 3a, label embeddings for five relations including *Explanation*, *Textual-Organization*, *Topic-Comment*, *Evaluation* and *Topic-Change* show such behavior. *Textual-Organization*, *Topic-Comment*, and *Topic-Change* are classes with a small amount of data and it is difficult to obtain good performance on these classes in a classification task. The reason for *Explanation* and *Evaluation* is not clear, and we leave it to future work.

4.5. Separate Experiments on PDTB Explicit and Implicit Relations

Previous studies (Demberg et al., 2019; Sanders et al., 2018) indicate that it is much easier to obtain consistent results on aligning PDTB explicit relations with relations from the other frameworks, while implicit relations are generally ambiguous and the consistency is much lower. Therefore, we conducted experiments on PDTB explicit and implicit relations separately. We use *LbEncRoberta* in the experiments. After the data preprocessing step outlined in section 4.1, 12 explicit relations and 14 implicit relations remain in the experiments.

| Data | Acc. | F1 | Label emb. |
|----------|--------------------|--------------------|--------------------|
| explicit | 88.98(\pm 0.41) | 79.19(\pm 0.64) | 99.15(\pm 0.60) |
| implicit | 56.05(\pm 0.56) | 40.56(\pm 0.81) | 82.21(\pm 0.85) |

Table 3: Results of experiments on PDTB explicit relations and implicit relations.

The classification results and label embedding learning results indicate that the learnt label embeddings for PDTB explicit relations are representative of the classes while the performance on implicit relations is sub-optimal.

4.6. Ablation Study

We choose *LbEncRoberta* and conduct ablation studies with PDTB explicit and implicit relations combined, similar to the experimental settings in Table 1. The impact of each loss can be seen in Table 4.

| Loss | Acc. | F1 | Label emb. |
|----------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| $-\mathcal{L}_{ICL}$ | 68.22(\pm 0.44) | 53.65(\pm 1.13) | 91.36(\pm 0.73) |
| $-\mathcal{L}_{LCL}$ | 65.02(\pm 0.47) | 51.23(\pm 1.62) | 80.37(\pm 1.42) |
| $-\mathcal{L}_{LEC}$ | 69.32(\pm 0.30) | 57.57(\pm 0.87) | 94.36(\pm 0.37) |
| $-\mathcal{L}_{ICE}$ | 69.88(\pm 0.09) | 56.94(\pm 0.36) | 90.79(\pm 0.76) |
| <i>Total</i> | 69.34(\pm 0.46) | 58.10(\pm 0.15) | 94.23(\pm 0.74) |

Table 4: Effect of each loss on model performance.

As shown, the label-centered contrastive loss (\mathcal{L}_{LCL}) is of paramount importance for the model’s performance, followed by the instance-centered contrastive loss (\mathcal{L}_{ICL}) and canonical cross-entropy loss (\mathcal{L}_{ICE}). This differs from the findings in Zhang et al. (2022b), where \mathcal{L}_{ICL} is the primary contributing factor to their results, indicating the distinct nature of our respective tasks. \mathcal{L}_{LEC} has some effect on F1 score of the classification task.

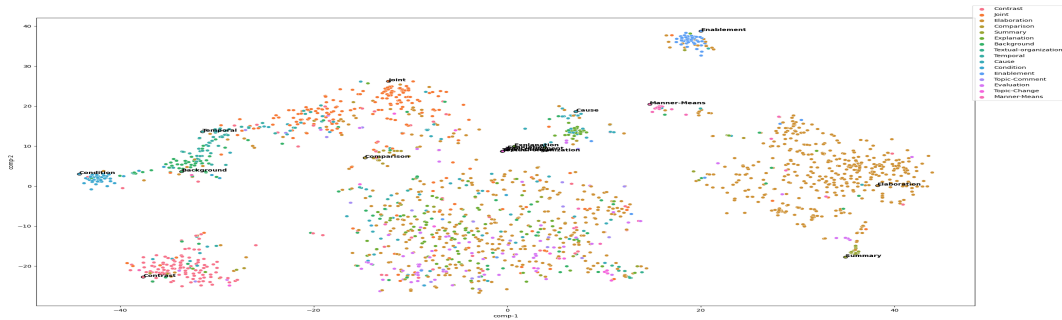
5. RST-PDTB Relation Mapping

5.1. Mapping Results

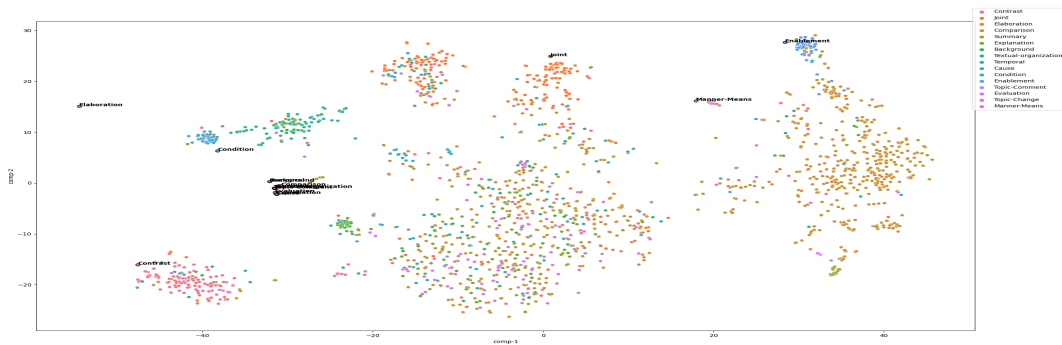
Table 5 shows the results of mapping 11 RST relations, with the five relations discussed in section 4.4 excluded, and 12 PDTB explicit relations discussed in section 4.5. Two relations with highest values in cosine similarity (greater than 0.10) are presented.

The table on the left shows the mapping results from RST’s perspective. For most of the RST relations, a PDTB relation can be identified as having a much higher value (≥ 0.40) than the others.

The table on the right shows the mapping results from PDTB’s perspective. As relation distributions are different, it is understandable that the two perspectives are not symmetric.



(a)



(b)

Figure 3: (a) Label embeddings learnt with data augmentation. (b) Label embeddings learnt without data augmentation. For visualization, we choose the label embeddings with the highest score from the three runs.

| RST | Relations in PDTB | PDTB | Relations in RST |
|--------------|--|-----------------|---|
| contrast | concession(0.25), contrast(0.24) | conjunction | contrast(0.22), elaboration(0.13) |
| manner-means | manner(0.30), purpose(0.25) | concession | contrast(0.25), elaboration(0.19) |
| cause | cause(0.40), level-of-detail(0.17) | cause | cause(0.40), manner-means(0.20) |
| background | synchronous(0.23), manner(0.16) | level-of-detail | manner-means(0.25), summary(0.23) |
| condition | condition(0.39), purpose(0.18) | synchronous | background(0.23), joint(0.20) |
| elaboration | concession(0.19), disjunction(0.14) | disjunction | joint(0.25), temporal(0.16) |
| enablement | manner(0.24), purpose(0.18) | manner | manner-means(0.30), enablement(0.24) |
| summary | contrast(0.35), level-of-detail(0.23) | condition | condition(0.39), summary(0.15) |
| joint | disjunction(0.25), synchronous(0.20) | substitution | manner-means(0.17), summary(0.17) |
| temporal | asynchronous(0.24), purpose(0.20) | asynchronous | temporal(0.24), joint(0.19) |
| comparison | purpose(0.17), level-of-detail(0.16) | contrast | summary(0.35), background(0.13) |
| | | purpose | manner-means(0.25), temporal(0.20) |

Table 5: Mapping between 11 RST relations and 12 PDTB explicit relations. The values in brackets represent the cosine similarity scores.

5.2. Extrinsic Evaluation

We compare our results with those provided by Costa et al. (2023), where the approach proposed in Demberg et al. (2019) is adopted but re-

sults are updated to PDTB 3.0. As shown in section 4.5, label embeddings learnt for PDTB explicit relations are more reliable and we choose to focus on the mapping between PDTB explicit relations and RST relations. Based on Table 5, we exclude PDTB’s *Substitution* relation in the experiments, for which no RST relations with higher similarity are observed, and relabel 11 PDTB explicit relations with RST labels based on Table 6.

While we choose the RST label mostly based on cosine similarity shown in Table 5, we take distribution of relations into account. For example, PDTB’s *Conjunction* relation is not mapped to RST’s *Contrast* relation but to *Elaboration*, because *Conjunction* is a large class in PDTB, similar to *Elaboration* in RST, and relabelling in this way may keep the label distribution balanced. Meanwhile, in our preliminary experiments, mapping PDTB’s *Contrast* relation to RST’s *Summary* relation yields poor performance. Therefore, we relabel PDTB’s *Contrast* as RST’s *Contrast* relation based on the results from RST’s perspective.

Similarly, we relabel PDTB explicit relations

based on the results shown in [Costa et al. \(2023\)](#)⁵. As their results are a mapping of 12 fine-grained RST relations and seven L2 PDTB relations, a direct mapping comparable to ours is not available. Thus, for a PDTB relation, if there are multiple mapped RST relations that fall under a broad class, the corresponding RST relation from the 16 categories is chosen, and the average of the percentages for the mapped classes is taken as the mapping strength, similar to cosine similarity in our results. For instance, PDTB *Concession* is mapped to *Contrast* (61.0%), *Antithesis* (84.0%), and *Concession* (88.0%), which are fine-grained relations under RST *Contrast*, and the mapping strength is the average of the three percentages, i.e., 0.78.

| Original PDTB —Sense Labels | RST Labels —Our method | RST Labels —Costa et al. (2023) |
|--------------------------------|---------------------------|------------------------------------|
| concession | contrast (0.25) | contrast (0.78) |
| contrast | contrast (0.24) | contrast (0.26) |
| conjunction | elaboration (0.13) | joint (0.84) |
| manner | manner-means (0.30) | — |
| cause | cause (0.40) | explanation (0.69) |
| synchronous | background (0.23) | temporal (0.98) |
| condition | condition (0.39) | condition (0.84) |
| disjunction | joint (0.25) | — |
| asynchronous | temporal (0.24) | temporal (0.94) |
| level-of-detail | manner-means (0.25) | — |
| purpose | manner-means (0.25) | — |

Table 6: Relabelling of PDTB explicit relations. The similarity scores are shown in brackets.

Based on our alignment results, 14964 instances of PDTB explicit relations are relabeled, and with the result in [Costa et al. \(2023\)](#), 13905 PDTB instances are relabeled. Adding PDTB data to RST data causes a marked performance drop. The best result is obtained with an ensemble model, which is formed by stacking a model trained with a target of minimizing supervised contrastive loss, a model trained to minimize a label embedding loss, the label embeddings being randomly initialized, and a model that takes the input for relation classification. The output distributions of the three models are averaged and used for model prediction, and a cross-entropy loss is to be reduced in addition to the supervised contrastive loss and label embedding loss. As shown in Table 7, the performance with our method is slightly higher.

| | Acc. | F1 |
|-------------------------------------|--------------|--------------|
| Costa et al. (2023) | 62.13 ± 0.34 | 46.96 ± 0.43 |
| Our method | 63.13 ± 1.12 | 47.95 ± 1.07 |
| -PDTB aug. | 63.82 ± 1.07 | 48.72 ± 0.11 |

Table 7: Results of extrinsic evaluation.

6. Conclusions

We propose a method of automatically aligning discourse relations from different frameworks. By em-

⁵Table 5 in their paper.

ploying label embeddings that are learned concurrently with input representations during a classification task, we are able to circumvent the challenges posed by segmentation differences, a significant hurdle encountered in prior studies. We perform intrinsic and extrinsic evaluation of the results of the method. Similar to other empirical studies, our method is affected by the amount of data, and we have to exclude some relations for which there may be too little training data to learn reliable label embeddings. A comparison with a theoretical proposal, such as ISO 24617-8 ([Prasad and Bunt, 2015](#)), merits investigation in future work. The method may extend beyond labelling of discourse relations to alignment of any label sets, leaving the possibility of application to a variety of scenarios.⁶

7. Acknowledgments

We thank the anonymous reviewers for insightful feedback and suggestions. Our thanks also go to Mark-Jan Nederhof for discussions and Craig Myles for the suggestion of using the diagonal entries of the normalized correlation matrix as a metric.

8. Bibliographical References

- Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2016. [Label-embedding for image classification](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1425–1438.
- Farah Benamara and Maite Taboada. 2015. [Mapping different rhetorical relation annotations: A proposal](#). In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.
- Peter Bourgonje and Olha Zolotareva. 2019. [Toward cross-theory discourse relation annotation](#). In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 7–11, Minneapolis, MN. Association for Computational Linguistics.
- Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. [Multi-view and multi-task training of RST discourse parsers](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1903–1913, Osaka, Japan. The COLING 2016 Organizing Committee.

⁶We thank the anonymous reviewers for pointing out the two directions.

- Harry Bunt and Rashmi Prasad. 2016. [ISO DR-Core \(ISO 24617-8\): Core concepts for the annotation of discourse relations](#). In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54.
- Lynn Carlson and Daniel Marcu. 2001. [Discourse tagging reference manual](#). *ISI Technical Report ISI-TR-545*, 54(2001):56.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. [Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory](#). In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.
- Soumya Chatterjee, Ayush Maheshwari, Ganesh Ramakrishnan, and Saketha Nath Jagarlapudi. 2021. [Joint learning of hyperbolic label embeddings for hierarchical multi-label classification](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2829–2841, Online. Association for Computational Linguistics.
- Christian Chiarcos. 2014. [Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4569–4577, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Nelson Filipe Costa, Nadia Sheikh, and Leila Kosseim. 2023. [Mapping explicit and implicit discourse relations between the RST-DT and the PDTB 3.0](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 344–352, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Vera Demberg, Merel CJ Scholman, and Fateh Torabi Asr. 2019. [How compatible are our discourse annotation frameworks? insights from mapping rst-dt and pdtb annotations](#). *Dialogue & Discourse*, 10(1):87–135.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yingxue Fu. 2022. [Towards unification of discourse annotation frameworks](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 132–142, Dublin, Ireland. Association for Computational Linguistics.
- Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. [Supervised contrastive learning for pre-trained language model fine-tuning](#). In *International Conference on Learning Representations*.
- Eduard H Hovy and Elisabeth Maier. 1992. [Par-simonious or profligate: how many and which discourse structure relations?](#) Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.
- Yin Jou Huang and Sadao Kurohashi. 2021. [Extractive summarization considering discourse and coreference relations based on heterogeneous graph](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3046–3052, Online. Association for Computational Linguistics.
- Yangfeng Ji and Jacob Eisenstein. 2014. [Representation learning for text-level discourse parsing](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.
- Yangfeng Ji and Jacob Eisenstein. 2015. [One Vector is Not Enough: Entity-Augmented Distributed Semantics for Discourse Relations](#). *Transactions of the Association for Computational Linguistics*, 3:329–344.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. [Supervised contrastive learning](#). *Advances in Neural Information Processing Systems*, 33:18661–18673.
- Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. [Implicit discourse relation classification: We need to talk about evaluation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.
- Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, and Bonnie Webber. 2006. Complexity of dependencies in discourse: are dependencies in discourse more complex than in syntax? In *5th International Workshop on Treebanks and Linguistic Theories*.

- Alan Lee, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2008. Departures from tree structures in discourse: Shared arguments in the penn discourse treebank. In *Proceedings of the constraints in discourse iii workshop*, pages 61–68.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [RoBERTa: A robustly optimized BERT pretraining approach](#).
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- William C Mann and Sandra A Thompson. 1988. [Rhetorical structure theory: Toward a functional theory of text organization](#). *Text*, 8(3):243–281.
- Daniel Marcu. 1996. [Building up rhetorical structure trees](#). In *Proceedings of the National Conference on Artificial Intelligence*, pages 1069–1074.
- Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT press.
- Taro Miyazaki, Kiminobu Makino, Yuka Takei, Hiroki Okamoto, and Jun Goto. 2019. [Label embedding using hierarchical structure of labels for Twitter classification](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6317–6322, Hong Kong, China. Association for Computational Linguistics.
- Johanna D. Moore and Martha E. Pollack. 1992. [A problem for RST: The need for multi-level discourse analysis](#). *Computational Linguistics*, 18(4):537–544.
- Karthik Narasimhan and Regina Barzilay. 2015. [Machine comprehension with discourse relations](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1253–1262, Beijing, China. Association for Computational Linguistics.
- Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. [Zero-shot learning with semantic output codes](#). In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). *Advances in Neural Information Processing Systems*, 32.
- R. Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind K. Joshi, Livio Robaldo, and Bonnie Lynn Webber. 2006. [The Penn Discourse Treebank 2.0 annotation manual](#).
- Rashmi Prasad and Harry Bunt. 2015. [Semantic relations in discourse: The current state of ISO 24617-8](#). In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. [The Penn Discourse TreeBank 2.0](#). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco. European Language Resources Association (ELRA).
- Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. [Discourse annotation in the PDTB: The next generation](#). In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. [Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1039–1046, Portorož, Slovenia. European Language Resources Association (ELRA).
- Ted JM Sanders, Vera Demberg, Jet Hoek, Merel CJ Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2018. [Unifying dimensions in coherence relations: How various annotation frameworks are related](#). *Corpus Linguistics and Linguistic Theory*.
- Tatjana Scheffler and Manfred Stede. 2016. [Mapping PDTB-style connective annotation to RST-style discourse annotation](#). In *Proceedings of the 13th Conference on Natural Language Processing*, pages 242–247.
- Manfred Stede, Stergos Afantenos, Andreas Peldszus, Nicholas Asher, and Jérémy Perret. 2016. [Parallel discourse annotations on a corpus of short texts](#). In *Proceedings of the Tenth International Conference on Language Resources and*

- Evaluation (LREC'16)*, pages 1051–1058, Portorož, Slovenia. European Language Resources Association (ELRA).
- Xu Sun, Bingzhen Wei, Xuancheng Ren, and Shuming Ma. 2017. [Label embedding network: Learning label representation for soft training of deep networks](#). *arXiv preprint arXiv:1710.10393*.
- Varsha Suresh and Desmond Ong. 2021. [Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Fei Wang, Yunfang Wu, and Likun Qiu. 2012. [Exploiting discourse relations for sentiment analysis](#). In *Proceedings of COLING 2012: Posters*, pages 1311–1320, Mumbai, India. The COLING 2012 Organizing Committee.
- Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. [Joint embedding of words and labels for text classification](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331, Melbourne, Australia. Association for Computational Linguistics.
- Bonnie Webber. 2004. [D-LTAG: extending lexicalized TAG to discourse](#). *Cognitive Science*, 28(5):751–779. 2003 Rumelhart Prize Special Issue Honoring Aravind K. Joshi.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, and Manabu Okumura. 2021. [Fusing label embedding into BERT: An efficient improvement for text classification](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1743–1750, Online. Association for Computational Linguistics.
- Honglun Zhang, Liqiang Xiao, Wenqing Chen, Yongkun Wang, and Yaohui Jin. 2018. [Multi-task label embedding for text classification](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4545–4553, Brussels, Belgium. Association for Computational Linguistics.
- Kun Zhang, Le Wu, Guangyi Lv, Enhong Chen, Shulan Ruan, Jing Liu, Zhiqiang Zhang, Jun Zhou, and Meng Wang. 2023. [Description-enhanced label embedding contrastive learning for text classification](#). *IEEE Transactions on Neural Networks and Learning Systems*.
- Oliver Zhang, Mike Wu, Jasmine Bayrooti, and Noah Goodman. 2021. [Temperature as uncertainty in contrastive learning](#). *arXiv preprint arXiv:2110.04403*.
- Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. 2022a. [Use all the labels: A hierarchical multi-label contrastive learning framework](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16660–16669.
- Zhenyu Zhang, Yuming Zhao, Meng Chen, and Xiaodong He. 2022b. [Label anchored contrastive learning for language understanding](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1449, Seattle, United States. Association for Computational Linguistics.

9. Language Resource References

- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. *RST Discourse Treebank*. distributed via LDC. Philadelphia: Linguistic Data Consortium: LDC2002T07, Text resources, 1.0, ISLRN: 299-735-991-930-2.
- Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. *Penn Discourse Treebank Version 3.0*. LDC. distributed via LDC. Philadelphia: Linguistic Data Consortium: LDC2019T05, Text resources, 3.0, ISLRN 977-491-842-427-0.

A New Annotation Scheme for the Semantics of Taste

Teresa Paccosi^{1,2}, Sara Tonelli²

¹ Università degli Studi di Trento, Italy

² Fondazione Bruno Kessler, Trento, Italy
tpaccosi@fbk.eu, satonelli@fbk.eu

Abstract

FrameNet serves as a comprehensive lexical database intended to represent contemporary language usage. However, it faces challenges in accurately representing specialized domains. Among these domains, FrameNet presents difficulties in capturing the specific semantics of human senses. Senses such as smell and taste are in fact included in more general frames or inadequately represented. Building on a previous resource proposing a new framework for olfactory events, we propose a similar annotation scheme for gustatory references in English, enlightening the potential of frames to effectively capture sensory semantics. Having a comprehensive framework to deal with the annotation of this kind of references in textual data is especially important to develop systems for the automatic extraction of sensory information. Moreover, our approach incorporates words from specific historical periods, thereby enriching the framework's utility for studying language in a diachronic perspective. In this paper, we introduce the annotation guidelines for taste and a preliminary annotation of culinary documents done using this approach.

Keywords: Sensory Language, Corpus Annotation, Frame Semantics, FrameNet

1. Introduction

The study of human senses is a fascinating topic that has always attracted scholars from different fields, such as philosophy, linguistics, cognitive science and neuroscience. Despite their importance, few works have dealt with the topic in the field of Natural Language Processing (NLP) or Digital Humanities (DH). Indeed, the development of automatic systems for the extraction of sensory-related information lacks a comprehensive linguistic framework to capture the semantics of specific senses. In this paper, we propose annotation guidelines for the semantics of taste in English, inspired by the annotation guidelines proposed for smell in [Tonelli and Menini \(2021\)](#). Our aim is to propose a comprehensive framework to capture the semantics of taste with a twofold purpose. On the one hand, we want to test whether a FrameNet-like approach, already proposed for smell, can be applied also to other senses, leading to the creation of comparable sensory benchmarks that can be used for different sensory studies. On the other hand, differently from FrameNet, we include in the lexical units for taste also obsolete terms in order to create resources that can be used for diachronic studies. Furthermore, the annotation guidelines proposed in this paper are specifically intended as a first step towards the development of an automatic system for the extraction of gustatory information for linguistic and historical studies.

Taste, together with smell, is especially interesting for its tendency to appear in emotionally charged contexts and to present a more evaluative content in its vocabulary. This tendency has shown that the most suitable way to study this sense is

by focusing not only on single words but also on their context ([Sneffjella and Kuperman, 2016](#); [Winter, 2019](#)), making frame semantics an appropriate framework for dealing with its study. In paragraph 3, we present the annotation guidelines for taste based on FrameNet, which entail a detailed examination of each label. Subsequently, we provide an overview of the annotation process with a preliminary annotation conducted on household and cooking recipe manuals, encompassing a temporal span of five centuries.

2. Related Work

Among the few works that have dealt with the topic of sensory language in NLP, [Tekiroğlu et al. \(2014\)](#) introduced Sensicon, a sensorial lexicon aiming to automatically associate English words with senses. This resource contained lemma-POS pairs with associated modality degrees for all five senses. Additionally, researchers have analyzed specialized lexicons used by reviewers to describe whisky and wine, focusing on taste and smell. [Hamilton and Lahne \(2020\)](#) developed a flavor wheel for whisky using a descriptive lexicon, while [Lefever et al. \(2018\)](#) aimed to predict wine characteristics from wine review corpora. The goal of these works is to identify words descriptive of perceptual experiences. Concerning taste specifically, there has been a growing interest in food representation, particularly for health-related studies. Some studies have focused on automatically extracting food entities, developing named-entity recognition (NER) models to support biomedical research and food science ([Cenikj et al., 2020](#); [Stojanov et al., 2021](#)). The authors constructed specialized corpora, pri-

marily emphasizing nutrient descriptions, quantities and food composition, by annotating recipes sourced from culinary social networks and websites (Popovski et al., 2019; Wróblewska et al., 2022). From a linguistic point of view, it has been recognized that understanding the semantics of human senses requires considering context, as already noted in Tonelli and Menini (2021), where the authors proposed an olfactory annotation framework based on FrameNet. Framing an entire event with its semantic roles enables a more holistic understanding of sensory information beyond isolated words. Their methodology was then used to create a multilingual benchmark (Menini et al., 2022), intended as a training for a supervised system for the automatic extraction of olfactory information which was used to analyse shifts in the perception of specific smell-related objects over time (Menini et al., 2023; Paccosi et al., 2023). The use of Frame Semantics to analyze taste-related language was successfully proposed in Diederich (2015). The author analyzed the use of two specific gustatory adjectives, *crispy* and *crunchy*, and the frames they trigger in both food experts' and everyday language. Through careful collocational analysis, the author elucidates the methodological strength of the frame-semantic approach in dealing with context analysis. By examining the evoked frames, the author demonstrates that even two adjectives considered synonyms can have different contexts of use upon thorough analysis. From a diachronic perspective, Bagli (2021) proposed an investigation into the vocabulary used to discuss gustatory experiences in English and the evolution of their semantic elaboration through the conceptual mechanisms of metaphor and metonymy. He argues that despite the disparagement that taste has undergone over time, it is a sense that has played an important role in shaping our conceptualization of emotions, with several metaphors based on its lexicon.

3. Annotation Guidelines for Gustatory Events

The present annotation guidelines for taste references in texts are inspired by the ones proposed for smell in Tonelli and Menini (2021). Their work puts its bases on the linguistic framework of frame semantics (Fillmore, 1976; Fillmore and Baker, 2001), implemented through the FrameNet annotation project (Ruppenhofer et al., 2006). The goal of FrameNet is to capture events and situations mentioned in texts. Frames represent constructs (i.e. events or situations) that function as the basis of our knowledge to understand the meaning of the words. For example, a word like the verb “talk” evokes an entire scenario implying at least two people in-

involved in a conversation. The events in FrameNet are modeled as a set of semantic roles or *frame elements* (FEs), which are typically the participants in the event, all connected to a *lexical unit* (LU) (i.e., the textual anchor that triggers the event or situation). In their work, Tonelli and Menini (2021) propose an adaptation of FrameNet to the olfactory domain, where only situations related to smell are annotated and specific semantic roles connected to olfactory events are identified. While FrameNet is a general-purpose framework including several frames to describe the perceptual experience of smell, the authors consider only one smell-related event that they call the *Olfactory Frame* (or *Olfactory Event*). They borrowed some general FEs from FrameNet and added some domain-specific ones that are self-explanatory and not ambiguous (e.g., *Evoked Odorant*, *Smell Source*) to facilitate a good agreement among annotators. As was done for smell, we define a single frame for taste: the *Gustatory Frame*. By searching for the lexical unit “taste” (both as a noun and as a verb) in FrameNet, we found 4 frames containing it: *Perception active*, *Sensation*, *Perception Experience* and a more specific one, presenting only two LUs (“taste” and “try”), *Tasting*. A part from them, we consider as taste-related frames also the *Ingestion* and the *Food* frames, from which we borrow some similar frame elements in our annotation guidelines. In the next sections, we present in more detail the two main components of the *Gustatory Frame*: LUs and FEs.

3.1. Lexical Units

In the choice of LUs for the *Gustatory Frame*, we select taste words incorporating lexical terms from different resources. The selected words have different part-of-speech including nouns, verbs, adverbs and adjectives, in line with FrameNet practice. The selection was conducted starting from the mental lexicon of De Deyne et al. (2019) and from WordNet (Miller, 1995). For the diachronic insights, we select lexical terms from the Historical Thesaurus of English (Kay, 2009)¹ from the “Taste/Flavour” category considering only those terms that are included in our temporal span (1500-2000). This combination of cognitive, contemporary, and historical lenses ensures that our selected LUs are both representative of current usage but also of the linguistic evolution of English, providing a robust foundation for our frame-based approach also in a diachronic perspective. A list of the LUs for taste is provided in Table 1.

¹<https://ht.ac.uk/>

| Lexical Units for Taste |
|---|
| Nouns: acidity, aftertaste, aroma, bitterness, dainty, delicacy, disgust, distaste, flavor, flavour, flavorful, flavourful, flavoring, flavouring, flavorsome, flavoursome, flavorful, flavourous, gustation, insipidity, mistaste, over-eating, palatableness, piquancy, pungency, rancidity, relish, relish (obsolete), saltiness, sapidity, sapor, savor, savoriness, savour, season, seasoning, sharpness, smack, smatch (obsolete), sourness, sowreness (obsolete), sweetness, tang, tarage, tartness, tast (obsolete), taste, tastelessness, tasting, unsavoriness, unsavouriness |
| Verbs: drink, drink up, eat, eat up, distaste, mistaste, partake, relish, season, smack, smatch (obsolete), sweeten, taste |
| Adjectives: acid, acidic, appetizing, appetizing, bitter, bitter-sweet, bland, dainty, delectable, delicious, delightsom(e), disgusting, flavorless, flavorful, flavourful, flavourless, flavoursome, gamy, indigestible, insipid, juicy, mellow, palatable, piquant, pungent, racy, rancid, rank, salt/salty, sapid, savory, savoury, savoury, seasoned, sharp, sour, soured, sower (archaic form of sour), spicy, stale, sweet, tangy, tart, tasteless, tasty, toothsome, unpalatable, unsaveor, unsavour, unsavoury, unsavory, unseasoned, unsweet, unsweetened, wearish, wersh (obsolete), yummy |
| Adverbs: sweetly, sourly, tastefully |

Table 1: List of Lexical Units for Taste

3.2. Frame Elements

The selected FEs encompass all potential participants contributing to frame activation along with lexical units. We first outline their differences or similarities with FrameNet’s FEs. Subsequently, we present more in detail each FE, providing some instances extracted from the annotated dataset. The FE *Taste Source* is a concept similar to the frame element *Food* for *Tasting* and *Ingestible* for *Ingestion* frames in FrameNet. In the same vein, *Quality* is a concept similar to the *Descriptor* FE in the context of food, and the semantic role of *Taster* aligns with *Agent* in *Tasting* and *Ingestor* in *Ingestion*. The *Location* FE of taste events finds a counterpart in the *Place* FE for both *Tasting* and *Ingestion* frames. There are no direct correspondence in FrameNet for several of the FEs contained in our framework, such as *Taste Modifier*, *Taste Carrier*, *Evoked Flavor*, *Circumstances*, and *Effect*, which we specifically created as domain-specific for the *Gustatory Frame*. Current FrameNet schema does not fully encapsulate aspects of the gustatory event that can be relevant for the study of sensory language. These domain-specific FEs could be viewed as extensions or specializations of existing ones, tailored to capture the unique semantic and experiential dimensions of taste. Our proposal aims at showing the relevance of FrameNet in capturing semantic content but also at underscoring the necessity for its continuous augmentation to accommodate the richness and specificity of human experience as captured through language. In the next sections, we present each FE in detail. In the example sentences, FEs will be represented between brackets, while the LUs are underlined. Taste frame elements have been defined with the goal to align with those for smell, facilitating comparison while emphasizing unique semantic structures. While certain elements such as *Effect* and *Location* remain identical, others such as *Evoked Taste* and *Taste Carrier* are complementary counterparts, with *Taste modifier* as the only label exclusive to gustation.

3.2.1. Taste Source

This FE corresponds to the person, animal or object that possesses a specific taste. It can refer to (non)human/object having a taste/flavor (e.g., plant, animal, perfume, human). This FE (between brackets) is the entity or object that the taster experiences through his or her senses. It is important to notice that if the taste source presents an adjective that describes it - see “slimy” in example 1 - this has to be annotated as part of the taste source. If the adjective refers instead to the perception of that specific source - see “unpleasant” in example 1 - then it has to be annotated as Quality of the LU:

1. [Slimy milk] has an unpleasant taste
2. When [the lettuce] is too young, the flavor is bitter

3.2.2. Taste Modifier

The object or animal that with its own taste/flavor can modify, alter or adding something different to the perception of the taste of a specific taste source. It is usually represented by ingredients that are added to a main course/food and often introduced by the verb “to season” and the preposition “with” followed by a noun. If there is more than one element, they have to be annotated as separate spans (see example 1):

1. Place two thick chops (of mutton) in a wooden dish and season lightly [with salt] and [pepper]
2. Factitious Port Wine is flavored [with a tincture drawn from the seeds of raisins]

3.2.3. Taste Carrier

This FE corresponds to the carrier of a taste, which can be a liquid such as water, spirits or liquors, or the container of the taste source (glass, plate, etc.). Note that the taste carrier has a different role both from Taste Source and Taste Modifier. The taste carrier is an object/person/animal which carries the taste of something else which is usually described

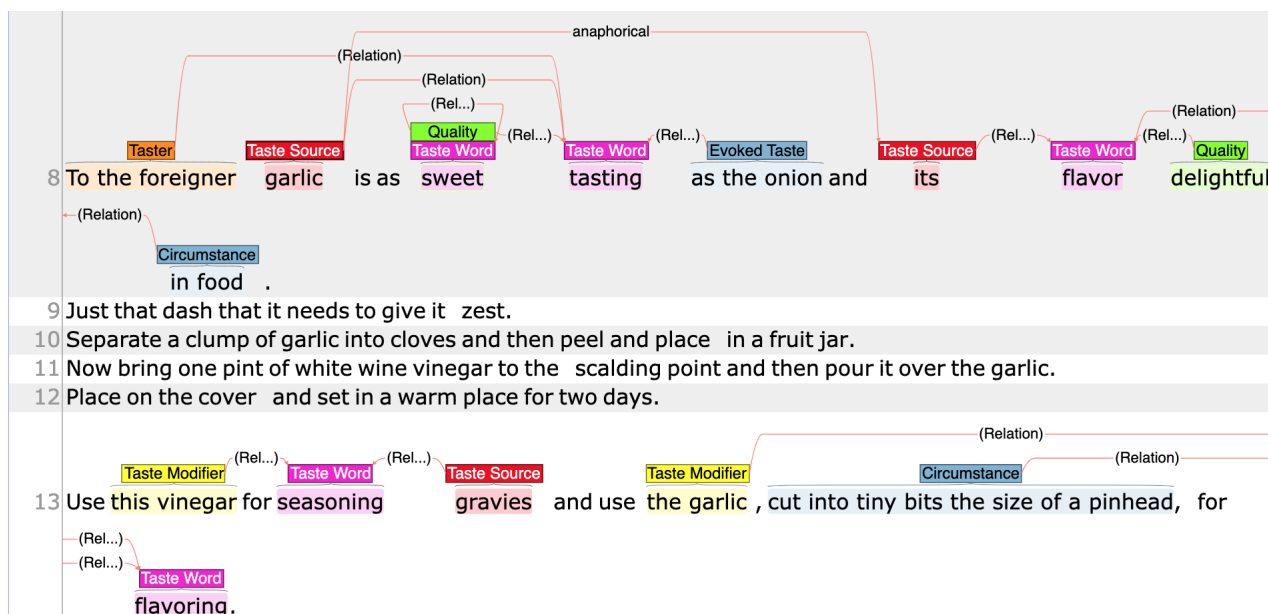


Figure 1: Screenshot of the INCEPTION tool used for taste annotation

as the object of perception by using the carrier. The distinction with taste modifier is important because there are few cases in which liquids/ointments are not modified by the taste of something but are the carriers of that taste. In example (1), “a considerable portion of essential oil” is the Taste Carrier, while the Taste Source is represented by “of the seeds”. Since sometimes this distinction is not clear, Taste Carrier should be annotated only when there is a clear distinction with the Taste Source and the Taste Modifier. This means that when a taste is described as coming generically from an object and it is not specified or clear from the context whether the object is the source or the container of the taste, the annotation as Taste Source should be preferred:

1. Only [a considerable portion of essential oil] has the flavour and taste of the seeds
2. Mr. Bland went into the hotel and drank [a glass or two] of wine and water

3.2.4. Quality

This is a quality associated with a taste and used to describe it. For example, sweet, disgusting, etc. This is typically expressed by qualitative adjectives. It is often preceded by an intensifier such as “very”, “really”. The intensifier has to be annotated with the related adjective in the same span, a part when the *Quality* is also a LU. In that case, the intensifier becomes a Quality of the Taste Word, with a double annotation which relates to itself (see Fig. 2). Qualities include intensity (weak, strong), duration (lingering, aftertaste), state (old, deteriorated),

character (quick, fruity), or hedonic characteristics (disgusting, pleasant, delicious). There are cases in which the Quality can also be a Taste Word, and has to be annotated with both the labels with the Quality FE linked to the Taste Word:

1. Cassia has a [slimy] [mucilaginous] taste
2. A taste which imparts a greater relish to the food is called [saline]



Figure 2: Double annotation of Taste Word and Quality

3.2.5. Taster

This FE refers to the human being/animal who perceives a taste with his/her mouth, has a perceptual experience of the taste. It is usually a personal pronoun, a possessive pronoun or a noun. The taster can also be expressed by mentioning the perceptive organ (e.g., palate, mouth) used in the gustatory experience (see example 2):

1. [To the foreigners] garlic is as sweet tasting as the onion
2. [Your palate] will reject them on account of their bitterness

| Century | 1500 | 1600 | 1700 | 1800 | 1900 |
|----------------|------|------|------|------|------|
| Document | 3 | 4 | 4 | 4 | 6 |
| Taste_Word | 212 | 381 | 169 | 376 | 278 |
| Taste_Source | 205 | 323 | 143 | 272 | 216 |
| Taste_Modifier | 130 | 102 | 60 | 89 | 68 |
| Quality | 83 | 204 | 60 | 171 | 128 |
| Taste_Carrier | 0 | 2 | 0 | 3 | 4 |
| Evoked_Taste | 1 | 2 | 3 | 12 | 2 |
| Location | 4 | 4 | 0 | 8 | 2 |
| Taster | 0 | 36 | 25 | 30 | 11 |
| Circumstance | 4 | 40 | 9 | 62 | 43 |
| Effect | 0 | 2 | 0 | 4 | 3 |

Table 2: Statistics of the Annotated Documents

3.2.6. Location

This FE describes the location/place where the taste event takes place, or the taste of a taste source is perceived. Locations can include both named places (for example names of cities), and common nouns describing physical locations such as garden, kitchen, room, etc.:

1. And [a neatly laid table] will stand before you with the most delicious food on it
2. He ordered the cat to be taken down [into the kitchen] and given something to eat and drink

3.2.7. Evoked Taste

This FE describes the person, animal, object's taste that is evoked/reminded by tasting a specific taste source, even if it is not visible/present in the scene. In English, this is often part of a comparison or similarity using the verb or noun "taste" and introduced by "like", "as" or the verb "to resemble". It is used to describe a taste that is perceived, referring to another:

1. (Jombo) in taste it [is like to an apple]
2. Burgundy pitch has a [terebinthinate] odour and taste

3.2.8. Circumstances

This FE characterizes the condition or circumstance in which the taste event occurs. This includes also temporal expressions, which describe a duration or a specific moment in which the taste event takes place. This FE can describe causal implications that lead to or influence the tasting experience. Circumstances are used to describe all that conditions in which the taste of a specific taste source can be altered or limited to a specific moment/event. It has to be distinguished by *Taste Modifier* that only considers the object/person/animal which modifies the taste of the Taste Source with its own:

1. If eaten [in excess, especially in an unripe or overripe state], fruits may occasion a disturbance of the stomach and bowels, often of a severe form
2. Tea and coffee also contain an astringent called tannin, which gives the peculiar bitter taste to the infusions [when steeped too long]

3.2.9. Effect

This FE describes an effect or reaction caused by the taste of a specific Taste Source. This can include entire sentences or clauses describing another event, that is not necessarily a taste event. This can include also the description of emotions triggered in the Taster by the taste event or anything that can effect him/her in some way:

1. If eaten in excess, especially in an unripe or overripe state, fruits [may occasion a disturbance of the stomach and bowels, often of a severe form]
2. By the process of cooking, agreeable flavors are developed [which stimulate the appetite and the flow of digestive fluids]

4. Annotation Process

4.1. Annotation Workflow

For taste annotation, we use INCEpTION (Klie et al., 2018), a web-based annotation tool, easily customizable both for labels and relations. We provide a screenshot of the tool in Figure 1. In annotating taste events we follow FrameNet established annotation methodology: we start by annotating a lexical unit in a sentence, and then we identify and connect the possible FEs participating in the gustatory event. In the provided example sentences for each label, frame elements can encompass single words or entire phrases. The annotated spans include articles for all frame elements, while for the LUs (Taste Words), only single terms are annotated,

without considering determiners. In FrameNet only the relation between lexical units and frame elements is considered. In our case, we annotate also the so-called “anaphorical relations”, similar to the smell annotation process described in [Tonelli and Menini \(2021\)](#). This integration captures FEs that link back to previously mentioned concepts or entities within the text. This is a relation especially useful at document level, since it allows us to identify also frame elements expressed with a pronoun having its antecedent lexically expressed in a different text passage.

4.2. Dataset

We manually annotated 21 manuals for household and culinary recipes published between 1575 and 1927 to test the suitability of the annotation framework with texts having a greater density of taste-related terms. These documents are taken from different publicly available historical and literary repositories:

- *Early English Books Online (EEBO)*,² a repository of documents published between 1470 and 1790 in different domains (literature, philosophy, politics, religion, geography, history, politics);
- *Project Gutenberg*,³ a digital archive compiled on a volunteer basis, containing different repositories, mainly in the literary domain;
- *medievalcookery.com*,⁴ a list of texts freely available online relating to medieval food and ancient cooking recipes;
- *foodsofengland.co.uk*,⁵ an online library which holds the complete texts of several cook books from 1390 to 1974;
- *Wikisource*,⁶ an online digital library of free-content textual sources managed by the Wikimedia Foundation.

In Table 2, we show the statistics of the annotated corpus divided per century. Two expert linguists, who were trained on the taste guidelines, annotated a total of three documents from different time periods (1670, 1720, and 1920) to assess Inter Annotator Agreement (IAA). The computation of Krippendorff’s alpha at a span-level ([Krippendorff,](#)

2011) resulted in an IAA score of 0.70, indicating a moderate level of agreement. While this suggests a reasonable level of consensus, there remains potential for improvement. Upon closer examination of the discrepancies, it was observed that there is a general agreement regarding the choice of labels. However, the disagreement arises from inconsistencies in the selection of spans, particularly in the exact number of tokens encompassed within those spans, as seen in the following instance, where the label *Taste_Source* is correct but the tokens were selected in a different way:

1. Boil **[your biggest skirrets]***Taste_Source* and season them with cinnamon and nutmeg
2. Boil your biggest **[skirrets]***Taste_Source* and season them with cinnamon and nutmeg

This observation has prompted us to refine our guidelines, placing more emphasis on defining the span selection process accurately. We released the annotated corpus at <https://github.com/dhfbk/Taste-Annotation>.

5. Conclusion and Future Directions

In this paper, we introduced a comprehensive annotation scheme for taste semantics. Our goal was to propose a framework for capturing taste-related information in textual data, serving as a benchmark for developing automated systems to extract gustatory-related information, especially intended for historical and linguistic studies. We tested the suitability of a previous approach for smell analysis, expanding the annotation guidelines to a different sense, and conducted preliminary annotations on a small set of taste-related documents. In the future, we plan to extend the annotation to additional documents to create a corpus containing sufficient information for building an automatic classifier for gustatory information extraction. This annotation scheme is also capable of capturing obsolete terms, making it suitable for annotating historical taste-related documents in English. Such a system can be used to analyze changes in sensory vocabularies over time, enabling diachronic analysis of the evolution of sensory semantic fields, a topic that has been hardly explored thus far.

6. Bibliographical References

- Marco Bagli. 2021. *Tastes we live by: The linguistic conceptualisation of taste in English*, volume 50. Walter de Gruyter GmbH & Co KG.
- Gjorgjina Cenikj, Gorjan Popovski, Riste Stojanov, Barbara Korousic Seljak, and Tome Eftimov.

²<https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/>

³<https://www.gutenberg.org/>

⁴<https://www.medievalcookery.com/etexts.html?England>

⁵<http://www.foodsofengland.co.uk/references.htm>

⁶https://en.wikisource.org/wiki/Main_Page

2020. Butter: Bidirectional lstm for food named-entity recognition.
- Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The “small world of words” english word association norms for over 12,000 cue words. *Behavior research methods*, 51:987–1006.
- Catherine Diederich. 2015. *Sensory adjectives in the discourse of food: A frame-semantic approach to language and perception*. John Benjamins Publishing Company.
- C. Fillmore. 1976. Frame semantics and the nature of language *. *Annals of the New York Academy of Sciences*, 280.
- Charles J Fillmore and Collin F Baker. 2001. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, volume 6.
- Leah M Hamilton and Jacob Lahne. 2020. Fast and automated sensory analysis: Using natural language processing for descriptive lexicon development. *Food Quality and Preference*, 83:103926.
- Christian J Kay. 2009. Jane roberts, michael samuels and irené wotherspoon. *The Historical Thesaurus of the Oxford English Dictionary*.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability.
- Els Lefever, Iris Hendrickx, Ilja Croijmans, Antal Van den Bosch, and Asifa Majid. 2018. Discovering the language of wine reviews: A text mining account. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3297–3302. LREC.
- Stefano Menini, Teresa Paccosi, Serra Sinem Tekiroğlu, and Sara Tonelli. 2023. Scent mining: Extracting olfactory events, smell sources and qualities. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 135–140.
- Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetöğlü, Ger Dijkstra, et al. 2022. A multilingual benchmark to capture olfactory situations over time. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 1–10.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- Teresa Paccosi, Stefano Menini, Elisa Leonardelli, Ilaria Barzon, and Sara Tonelli. 2023. Scent and sensibility: Perception shifts in the olfactory domain. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 143–152.
- Gorjan Popovski, Barbara Koroušić Seljak, and Tome Eftimov. 2019. Foodbase corpus: a new resource of annotated food entities. *Database*, 2019:baz121.
- Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. [Framenet ii: Extended theory and practice](#). Working paper, International Computer Science Institute, Berkeley, CA.
- Bryor Sneffjella and Victor Kuperman. 2016. It’s all in the delivery: Effects of context valence, arousal, and concreteness on visual word processing. *Cognition*, 156:135–146.
- Riste Stojanov, Gorjan Popovski, Gjorgjina Genikj, Barbara Koroušić Seljak, and Tome Eftimov. 2021. A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: Algorithm development and validation. *Journal of Medical Internet Research*, 23(8):e28229.
- Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2014. Sensicon: An automatically constructed sensorial lexicon. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1511–1521.
- Sara Tonelli and Stefano Menini. 2021. [FrameNet-like annotation of olfactory information in texts](#). In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 11–20, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.
- Bodo Winter. 2019. *Sensory linguistics*. John Benjamins Publishing Company.

Ania Wróblewska, Agnieszka Kaliska, Maciej Pawłowski, Dawid Wiśniewski, Witold Sosnowski, and Agnieszka Ławrynowicz. 2022. Tasteset–recipe dataset and food entities recognition benchmark. *arXiv preprint arXiv:2204.07775*.

What to Annotate: Retrieving Lexical Markers of Conspiracy Discourse from an Italian-English Corpus of Telegram Data

Costanza Marini, Elisabetta Jezek

University of Pavia, Department of Humanities
Corso Strada Nuova, 65, 27100 Pavia, Italy
costanza.marini@unipv.it, elisabetta.jezek@unipv.it

Abstract

In this age of social media, Conspiracy Theories (CTs) have become an issue that can no longer be ignored. After providing an overview of CT literature and corpus studies, we describe the creation of a 40,000-token English-Italian bilingual corpus of conspiracy-oriented Telegram comments – the *Complotto* corpus – and the linguistic analysis we performed using the *Sketch Engine* online platform (Kilgarriff et al., 2010) on our annotated data to identify statistically relevant linguistic markers of CT discourse. Thanks to the platform’s *keywords* and key *terms* extraction functions, we were able to assess the statistical significance of the following lexical and semantic phenomena, both cross-linguistically and cross-CT, namely: (1) evidentiality and epistemic modality markers; (2) *debunking* vocabulary referring to another version of the truth lying behind the official one; (3) the conceptual metaphor INSTITUTIONS ARE ABUSERS. All these features qualify as markers of CT discourse and have the potential to be effectively used for future semantic annotation tasks to develop automatic systems for CT identification.

Keywords: conspiracy theories, corpus annotation, linguistic analysis

1. Introduction

Conspiracy Theories (CTs) are “allegation[s] of conspiracy that may or may not be true” (Douglas et al, 2019, p. 4), but whose proliferation in this age of social media poses a threat to society, in that they are contributing to distorting our perception of reality.

The goal of this study is to verify whether CTs are indeed characterised - as certain studies seem to suggest, albeit from a mostly monolingual perspective focusing on single CTs - by common discourse features (especially lexical and semantic), that may be exploited in future annotation tasks to develop automatic systems for CT detection.

In this contribution, after an interdisciplinary overview of CT literature, we describe the collection and annotation of our dataset and the creation of the *Complotto* corpus, an English-Italian bilingual corpus of conspiracy-oriented Telegram comments. The corpus counts 658 comments (317 Italian and 341 English), for a total of 40,045 tokens. For both English and Italian, comments were taken from three language-specific Telegram channels focusing on the same three CTs: Flat Earth, the vaccine conspiracy, and the climate change hoax.

The linguistic analysis we performed using the *Sketch Engine* online platform (Kilgarriff et al., 2010) on our annotated data allowed us to identify statistically relevant linguistic markers of CT discourse. In particular, thanks to the platform’s *keywords* and key *terms* extraction functions, we were able to determine the statistical significance of the following phenomena, both cross-linguistically and cross-CT, namely: (1) evidentiality and epistemic modality markers; (2) *debunking* vocabulary referring to another version of the truth lying behind the official one; (3) the

conceptual metaphor INSTITUTIONS ARE ABUSERS.

2. Conspiracy Theories

While the term *conspiracy* refers to an actual plot orchestrated at somebody’s expense, a CT refers to “an allegation of conspiracy that may or may not be true” (Douglas et al., 2019, p. 4). An example of conspiracy gone awry is the infamous Gunpowder Plot that took place on the 5th of November 1605, when a group of English Catholics, amongst which Guy Fawkes, attempted to blow up the House of Lords and to kill King James I. On the other hand, countless individuals and organisations have been accused of the assassination of U.S. President J.F. Kennedy in 1963, but all these CTs remain unproven.

Depending on the CT, the group of scheming individuals allegedly behind the plot might be identified with the same “epistemological authorities” (Uscinski, 2020) often tasked with making hard and unpopular decisions – i.e., scientists, governments, and other societal institutions.

As pointed out by Mancosu and Vassallo (2022, p. 2), CTs are often linked to a degeneration of public debate, especially in these times of social media, as in the case of the supposed electoral fraud to the detriment of Donald Trump which led to the assault on the U.S. Capitol Hill in January 2020.

By definition, CTs attempt to provide alternative explanations to events, therefore spreading the idea that “things are not as they seem, and that the truth behind certain events is hidden from view” (Demata et al., 2022, p.1) providing false evidence to support their claims (Danesi, 2023, p. 13).

As a consequence, CT narratives tend to juxtapose an Insider vs. an Outsider group (Bodner et al., 2020), where the Insider group feels threatened by the Outsider group and has to come up with strategies to counter these threats (Tangherlini, 2018).

In terms of the reasons why people believe in CTs, studies such as Douglas et al. (2017) have shown it satisfies a set of psychological motives, such as the desire for certainty, control, and security. Among the cognitive processes linked to a CT mentality, we find supernatural beliefs, a quasi-religious mentality, feelings of powerlessness and low control in the socio-political domain (Douglas, 2019, p. 7-9).

When it comes to analysing the CT discourse, we can expect all these characteristics, traits, and recurring motives to have their linguistic counterparts.

3. Related Works

Prior studies have focused on compiling CT corpora using both printed documents and social media content, mainly for English. For instance, Uscinski et al. (2011) compiled a corpus of conspiracy letters to the editor of *The New York Times* published from 1897 to 2010. On the other hand, Catenaccio (2022), who also carried out a corpus-driven analysis searching for the linguistic features of CT discourse, compiled a corpus of published books providing alternative accounts of the 9/11 events. Miani et al. (2021) released the *Language of Conspiracy* corpus (LOCO), an 88-million-word corpus of online texts covering a wide range of CTs collected automatically using a seeding approach. The LOCO corpus was used successfully by Mompelat et al. (2022) to design an annotation scheme that was used to develop CT vs. mainstream automatic document retrieval methods. Lastly, Russo et al. (2023) have created a dataset of 25.000 Italian posts extracted from five conspiracy-oriented Telegram channels, that were annotated to perform two computational classification shared tasks: a binary task aimed at determining whether a post is “conspiratorial” or not and a multi-class task aimed at recognizing the specific CT talked about in the post (Covid-19, QAnon, Flat Earth, Pro-Russia).

Previous works aimed at characterising the language of CTs have identified the following indicators: (1) constant reference to insider group vs. outsider group (Holur et al., 2022); (2) a non-standard use of epistemic stance and evidentiality markers (Catenaccio, 2022; Scharloth et al., 2019); (3) a creative *debunking* vocabulary referring to the fact that another version of the truth lies behind the official one (Ebling et al., 2013); (4) an instrumental use of conceptual metaphors to convey conspiratorial content (Danesi, 2023), since conceptual metaphors facilitate processing of non-literal meaning

allowing to view abstract concepts in terms of the properties of more concrete ones (Lakoff and Johnson, 1980).

Since no comprehensive cross-linguistic study has been conducted yet to define a shared annotation scheme for CT language, our work wishes to provide its contribution drawing from the afore-mentioned literature, as well as from existing ISO annotation standards that have not been applied to CT discourse yet.

4. Corpus Design and Annotation

In light of the relevance of social media in the spread of CTs and inspired by Russo et al. (2023), for our study, we decided to focus on Telegram data. If Twitter comments are indeed often short and very contextualised within a thread (Mompelat et al., 2022, p. 12), Telegram comments are usually quite long and exhaustive since they are not posted on a public wall but on the community’s channel, which works as a chatgroup and mainly includes ingroup users who wish to be kept informed.

We first compiled a dataset of Telegram posts from six openly CT-oriented channels fostering conversation on three different CTs: Flat Earth, the vaccine conspiracy, and the climate change hoax.

For each language, we found a different channel dedicated to the above-mentioned CTs, namely: *The Flat Earth Reality, No Vaccination. My Body My Choice* and *Climate Change HOAX* for English, and *Terra Piatta* ‘Flat Earth’, *Vax: le cavie siamo noi?* ‘Vax: are we the guinea pigs?’ and *Scie chimiche e clima* ‘Chemtrails and climate’ for Italian.

4.1 Annotation Tool

To obtain an actual CT corpus, we decided to trim out non conspiratorial comments by performing an annotation exercise. Two human annotators were asked to classify posts as either [Conspiratorial] or [Non-conspiratorial] using the *Taguette*¹ tool, a user-friendly and open-source online annotation environment. Other annotation tools were considered (e.g., INCEPTION, MAXQDA), but *Taguette* was eventually chosen because it allows to: import several file formats (including HTML); export the whole project, the annotated documents or just the performed annotation depending on the project’s needs; work remotely on the *Taguette* server; annotate by simply selecting the desired span of text and highlighting it using the desired labels.

We decided to import HTML documents in order to obtain an annotation-friendly visualisation of the Telegram data that preserves much of the original layout, thus enabling annotators to understand the chat dynamics. The files were first extracted from the selected Telegram channels,

¹ <https://www.taguette.org/>

slightly simplified using a clean-up tool² and then uploaded onto *Taguette* (Figure 1).

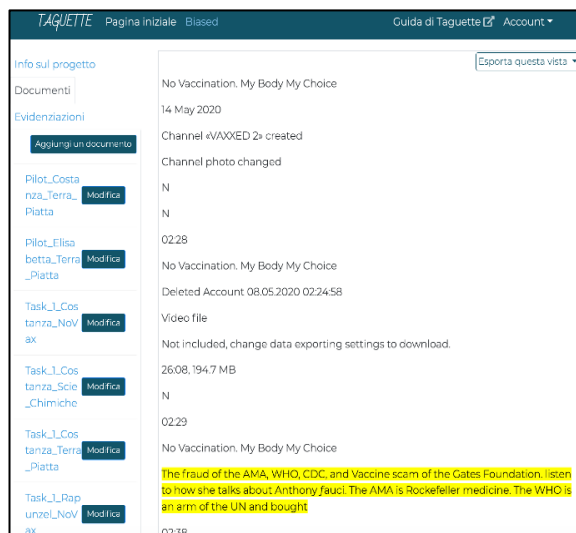


Figure 1: The *Taguette* annotation environment

The menu window on the left allows to access the project's info, the uploaded documents, and the existing annotations, called *evidenziazioni* 'highlights' because they appear as yellow highlights. On the right, you can see the beginning of the document *No vaccination. My Body My Choice*. Only one comment was annotated, namely a conspiratorial one: *The fraud of the AMA, WHO, CDC, and Vaccine scam of the Gates Foundation. listen to how she talks about Anthony Fauci. The AMA is Rockefeller medicine. The WHO is an arm of the UN and bought*.

4.2 Annotation Scheme and Guidelines

Dedicated guidelines were created to provide the two annotators³ with a general description of what CTs are (see § 2) so that they could distinguish [Conspiratorial] comments from [Non-conspiratorial] ones. It was chosen not to provide any linguistic or textual cues, in order not to skew the annotation results.

According to our annotation scheme, [Conspiratorial] comments are defined as "comments in which users (directly or indirectly) express themselves in favour of a CT", whereas [Non-conspiratorial] comments are "comments in which users (1) talk about CTs without expressing their stance or (2) talk about other topics, even unrelated ones".

The guidelines also specified what not to annotate, i.e., usernames, dates, times, Telegram channel names in isolation, footprints of multimedia files that were not included in the download, recurring comments that were not

actually written with a communicative aim but automatically posted in the chatroom. Moreover, what counted as comment was clearly specified. For instance, if a user conveyed content over several separately sent messages, each message was annotated as a separate comment. It was also agreed that the whole comment would be tagged and not smaller spans of text.

Figure 2 shows a portion of the chat conversation from the Telegram channel *The Flat Earth Reality*.

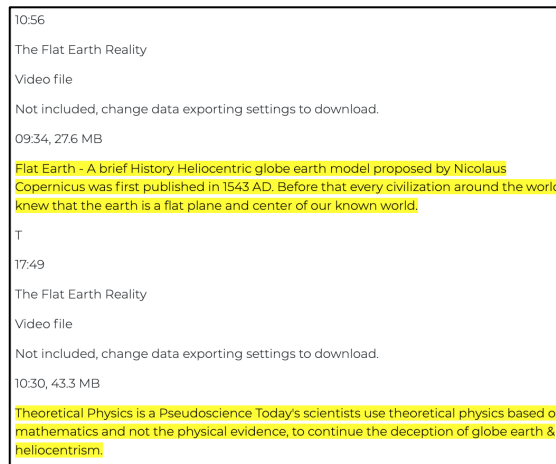


Figure 2: Annotating *The Flat Earth Reality* data

As you can see, the name of the channel, as well as dates, and multimedia files fingerprints were not tagged, while two comments were annotated as [Conspiratorial] because they both clearly show the author's belief that the Earth is flat and that the truth is being covered up (e.g., *scientists use theoretical physics based on mathematics and not the physical evidence, to continue the deception of globe earth & heliocentrism*).

4.3 Annotation Results

After a successful pilot test, the two annotators were asked to annotate at least 111 [Conspiratorial] comments per Telegram channel, in order to reach 1000 comments if the corpus was expanded to a third language⁴ following the same design, i.e., adding three other Telegram channels. By the end of the annotation, 1025 comments were annotated, of which 658 were identified as [Conspiratorial] and 304 as [Non-conspiratorial] by both annotators. According to Landis & Koch's (1977) interpretation of Cohen's κ , the two annotators reached a perfect level of Inter Annotator Agreement (IAA) in all documents, as summarised in Table 1 below.

² <https://www.htmlwasher.com/>

³ The two annotators are the first author of the paper and an archival specialist with prior experience in annotation.

⁴ The creation of a *German Complotto* corpus is currently underway.

| Telegram channel | 1st comment | Cohen's κ |
|---------------------------------|-------------|------------------|
| <i>Terra Piatta</i> | Nov 2016 | 0.86 |
| <i>Vax: le cavie siamo noi?</i> | July 2017 | 0.81 |
| <i>Scie chimiche e clima</i> | Nov 2022 | 0.86 |
| <i>The Flat Earth Reality</i> | Oct 2020 | 0.83 |
| <i>No Vaccination.</i> | May 2020 | 0.84 |
| <i>My Body My Choice</i> | Aug 2021 | 0.87 |

Table 1: IAA when annotating for [Conspiratorial] vs. [Non conspiratorial] comments.

Only the 658 comments that both annotators annotated as [Conspiratorial] – 317 Italian and 341 English – were included in the final corpus, which was then uploaded onto the *Sketch Engine* (Kilgarrieff et al., 2010) online platform to identify relevant indicators of CT discourse.

5. Corpus Methods and Results

The *Complotto* corpus counts 40,045 tokens but was uploaded onto the *Sketch Engine* platform as two separate corpora - the *Italian Complotto* and the *English Complotto* - of three documents each, in order to retrieve relevant language-specific and CT-specific features.

The *Italian Complotto* corpus counts a total of 22,252 tokens, of which 7,453 tokens in the *No-Vax* subcorpus (short for *Vax: le cavie siamo noi?*), 12,331 in the *Scie Chimiche* subcorpus and 2,468 in the *Terra Piatta* subcorpus. On the other hand, the *English Complotto* corpus counts 17,793 tokens, of which 4,612 in the *Climate Change* subcorpus, 8,434 in the *Flat Earth* subcorpus and 4,747 in the *My Body My Choice* subcorpus.

The two *Sketch Engine* functions that were used so far for the linguistic analysis are terminology extraction-related and are called *keywords* and *terms* (i.e., key multi-word expressions). They are able to extract keywords and terms by comparing the observed frequency data of a focus corpus and those of a larger reference corpus. Only words and multi-word expressions that appear with a statistically significant higher frequency in the focus corpus than in the reference one obtain *key status*. In our case, the reference corpus used for the *Italian Complotto* corpus was the *Italian Web 2020*, a 12-billion-word corpus made of Italian texts collected from the web, while the English reference corpus for the *English Complotto* was the *English Web 2021*, a 52-billion-word corpus of English web-crawled texts. As an advanced setting, we specified that keywords and terms should not be either too rare or too common and appear at least 4 times in the focus corpus to be considered for key status.

For the purposes of this study, we excluded CT-specific lexis from the analysis, such as the keywords *chemtrails* for the *Climate Change* subcorpus or *Flat Earth* for the *Flat Earth* one, because we are looking at common markers of CT discourse.

5.1 Italian Results

For the *Italian Complotto* corpus, *Sketch Engine*'s wordlist function was useful to confirm the insider vs. outsider group dynamics of the corpus (Holur et al. 2022), since the pronouns *noi* 'we' and *ci* 'us' appear as the most frequent personal pronouns overall.

Among the top five keywords of the Italian *No-Vax* subcorpus, we unsurprisingly found the lemmas *vaccino* 'vaccine', *vaccinare* 'to vaccinate', and *vaccinazione* 'vaccination', which were excluded from the analysis since they are CT-related. However, when looking at their observed modifiers, we found the expression *vaccinazione coercitiva* 'enforced vaccination', which contributes to indirectly conveying an interesting and novel conceptual metaphor, i.e., INSTITUTIONS ARE ABUSERS (specifically governments), which is in line with Lakoff and Johnson (1980)'s framework where an abstract target domain (as, in our case, INSTITUTIONS) is viewed in terms of a more concrete source domain (ABUSERS).

Among the top 50 keywords in the subcorpus we found three epistemically charged lemmas – *naturalmente* 'naturally', *assolutamente* 'absolutely' and *probabilmente* 'probably' – and seven lemmas clearly linked to the evidential necessity of providing a source for one's claims, i.e., *filmato* 'video', *news* 'news', *dichiarazione* 'declaration', *documentazione* 'documentation', *notizia* 'piece of news', *indicazione* 'direction', and *fonte* 'source'. These findings are in line with Scharloth et al. (2019) and Catenaccio (2022) and strongly suggest that the lexicon of the semantic field of EVIDENCE should be included among potential indicators of CT discourse. To introduce said evidence, we often find the presentative discourse marker *ecco* 'here' (e.g., *Ecco qui le prove schiaccianti... che non lasciano dubbi* 'Here the undeniable proof... which does not leave room for doubt'), which is also among the top 50 keywords of the subcorpus.

As a presentative discourse marker, *ecco* 'here' can be considered an indicator of dialogue acts (Bunt et al., 2010) characterised by an information-providing communicative function (ISO 24617-2), which seem particularly frequent in the whole corpus.

The only key multi-word term found in the subcorpus is *libertà di scelta* 'freedom of choice', which sums up the stance of chat members on the topic of vaccination and helps characterise them as the allegedly oppressed and threatened ingroup (Holur et al., 2022).

On the other hand, the most relevant *debunking* lemmas among the top 50 keywords in the *Scie Chimiche* subcorpus are the lemmas *mentire*, 'to lie', *menzogna* 'lie', and *manipolazione* 'manipulation', which all hint at the fact that another version of the truth allegedly exists but is hidden by official institutions (Ebling et al., 2013), in line with the conceptual metaphor

INSTITUTIONS ARE ABUSERS, specifically LIARS. No interesting key multi-word terms were found.

Finally, also among the top 50 keywords of the Italian *Terra Piatta* subcorpus, we find the lemmas *foto* 'photo', *test* 'test', *video* 'video', *prova* 'proof' – all belonging to the semantic field of EVIDENCE – as well as the lemma *fotomontaggio* 'fake photo' (actually among the top five), which implies that the evidence provided by scientific institutions such as NASA is actually unreliable and that, once again, INSTITUTIONS ARE ABUSERS, specifically LIARS.

5.2 English Results

Among the top 50 keywords of the *Climate Change* subcorpus, we can easily spot *debunking* expressions (Ebling et al., 2013) hinting at another version of the truth, such as *whistleblower* and *reportedly*. The latter is both an evidential and an epistemic marker used to signal that who writes is not the source of the information and that the information itself is not necessarily trustworthy (e.g., *Watch Argentinian Engineer Juan Baigorri Velar Reportedly Invented a Functional Rainmaker, But It Is Lost to History*). Similarly, the presence of the adjective *so-called*, which is always to be found within the key multi-word expression *so-called expert*, is aimed at discrediting the scientific community. The keyword *proof* fits among the lemmas making up the semantic field of EVIDENCE, together with the multi-word expression *full interview*. The presence of *poison* among the top keywords, and of *spray pollution* among the top terms, are both clear indicators of distrust in the institutions.

As for the *Flat Earth* subcorpus, the most interesting *debunking* keywords are *deception*, *hoax*, *debunk*, *scientism*, *conspiracy* and *fake*, which are all hinting at another alleged version of the truth (Ebling et al., 2013), as well as at the institutional and scientific responsibility behind said deception. This supports our proposed conceptual metaphor INSTITUTIONS ARE ABUSERS, specifically DECEIVERS. Last but not least, among the top fifty keywords of the *My Body My Choice* subcorpus, we can find the words *false*, *fully*, and *completely* - which are all "epistemically charged" expressions, to quote Catenaccio (2022, p. 31).

The next step in our analysis will be focused on corpus wordlists. From a first analysis of the *English Complotto* wordlist, we noticed several grammatical words that do not appear among the top 50 corpus keywords but have a considerably high rank in the *English Complotto* corpus and a much lower one in the *English Web 2021*. This is the case, for instance, of *here* (42nd most frequent word in the *English Complotto*, 132nd in the *English Web*) and *how* (54th in the *English Complotto*, 91st in the *English Web*). The following examples wish to provide insight on their use within the *English Complotto* corpus:

(1) Here's *why pilots can't prove curvature by demonstrating any change in level when flying between two locations, no matter how far apart.*

(2) *How can we be certain?*

(3) *Most people do not realise how integral artistic rendition is part of NASA's deception.*

Once again, the presentative discourse marker *here* in (1), and the conjunction *how* in (3) can be seen as indicators of information-providing communicative functions, while the use of *how* as an interrogative adverb in (2) points towards an information-seeking communicative function (Bunt et al., 2010).

6. Conclusion

According to the relevant literature, Conspiracy Theories (CTs) can be defined as "allegation[s] of conspiracy that may or may not be true" (Douglas et al, 2019, p. 4). However, their proliferation on social media is an undeniable threat to democratic societies.

In this study, after offering an interdisciplinary review of CT literature (§ 2) and related corpus works (§ 3), we have provided a detailed description of the design and annotation task that led to the creation of the *Complotto* corpus, a 40,000-token English-Italian corpus of conspiracy-oriented Telegram comments (§ 4). The *Complotto* corpus is a collection of 658 comments (317 Italian and 341 English) that were taken from three English- and three Italian-speaking Telegram channels focusing on the same three CTs: Flat Earth, the vaccine conspiracy, and the climate change hoax. In section 5, our corpus methods and results are explained.

The linguistic analysis we performed using the *Sketch Engine* online platform (Kilgarriff et al., 2010) on our annotated data allowed us to assess the statistical relevance (both cross-linguistically and cross-CT) of the following phenomena thanks to the platform's *keywords* and *key terms* extraction functions, namely: (1) evidentiality and epistemic modality markers; (2) debunking vocabulary conveying the idea that another version of the truth lies behind the official one; (3) the conceptual metaphor INSTITUTIONS ARE ABUSERS.

All these features qualify as markers of CT discourse and have the potential to be effectively used as tags for future fine-grained semantic annotation tasks to develop systems of automatic CT identification. On-going work is focused on adding a third language to the corpus, namely German.

7. Acknowledgments

We thank the second annotator that contributed to the creation of the *Complotto* corpus and the two reviewers for their useful comments.

8. Bibliographical References

- Bodner, J., Welch, W., and Brodie, I. (2020). *COVID-19 conspiracy theories: QAnon, 5G, the New World Order and other viral ideas*. McFarland, Jefferson.
- Bunt, H., Alexandersson, J., Carletta, J., Chloe, J.W., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., and Traum, D. (2010). Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, La Valletta, Malta, May. European Language Resource Association (ELRA).
- Catenaccio, P. (2022). A corpus-driven exploration of conspiracy theorizing as a discourse type. In M. Demata, V. Zorzi, & A. Zottola (Eds.), *Conspiracy Theory Discourses*. Amsterdam/ Philadelphia: John Benjamins Publishing Company, pp. 25–47.
- Danesi, M. (2023). *Politics, lies and conspiracy theories: A cognitive linguistic perspective*. Routledge, Oxon/ New York.
- Demata, M., Zorzi, V., and Zottola, A. (2022). Conspiracy theory discourses – Critical inquiries into the language of anti-science, post-truthism, mis/disinformation and alternative media. In M. Demata, V. Zorzi, & A. Zottola (Eds.), *Conspiracy Theory Discourses*. Amsterdam/ Philadelphia: John Benjamins Publishing Company, pp. 1–22.
- Douglas, K. M., and Leite, A. C. (2017). Suspicion in the workplace: Organizational conspiracy theories and work-related outcomes. *British Journal of Psychology*.
- Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., and Deravi, F. (2019). Understanding conspiracy theories. *Advances in Political Psychology*, 40(1):3–35.
- Ebling, S., Scharloth, J., Dussa, T., and Bubenhofer, N. (2013). Gibt es eine Sprache des politischen Extremismus? In *Die da oben. Sprache, Politik, Partizipation*. Bremen: Hempen Verlag, pp. 43–69.
- ISO. (2020). *Language resource management-Semantic annotation framework (SemAF) - Part 2 - Dialogue acts*. Standard, Geneva.
- Kilgarriff, A., Vitek, B., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.
- Lakoff, G., and Johnson. M. (1980). *Metaphors We Live By*. University of Chicago Press, Chicago.
- Landis, J.R., and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Mancosu, M., and Vassallo, S. (2022). The life cycle of conspiracy theories: evidence from a long-term panel survey on conspiracy beliefs in Italy. *Italian Political Science Review/ Rivista Italiana di Scienza Politica*, 52:1–17.
- Miani, A., Hills, T., and Bangerter, A. (2021). LOCO: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*.
- Mompelat, L., Tian, Z., Kessler, A., Luetgen, M., Rajanala, A., Kübler, S., and Seelig, M. (2022). How “Loco” Is the LOCO Corpus? Annotating the Language of Conspiracy Theories. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 111–119, Marseille, France. European Language Resources Association.
- Russo, G., Stoehr, N., and Horta Ribeiro, M. (2023). ACTI at EVALITA 2023: Automatic Conspiracy Theory Identification Task Overview. In *EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, Parma, Italy, September. CEUR.
- Scharloth, J., Obert, J., and Keilholz, F. (2019). Epistemische Positionierung in verschwörungstheoretischen Texten. Korpuspragmatische Untersuchung von epistemischer Modalität und Evidentialität am Beispiel der Holocaustleugnung. *Zeitschrift für Diskursforschung*, 4:159–198.
- Tangherlini, T.R. (2018). Toward a generative model of legend: Pizzas, bridges, vaccines, and witches. *Humanities*, 7(1):1.
- Uscinski, J. E., Parent, J., and Torres, B. (2011). Conspiracy theories are for losers. In *APSA 2011 Annual Meeting*.
- Uscinski, J. E. (2020). *Conspiracy theories. A Primer*. Rowman & Littlefield, London.

Lightweight Connective Detection Using Gradient Boosting

Mustafa Erolcan Er¹, Murathan Kurfalı², Deniz Zeyrek¹

¹Cognitive Science Dept., Graduate School of Informatics, Middle East Technical University

²Sensory-Cognitive Interaction Lab, Department of Psychology, Stockholm University
erolcan@metu.edu.tr, murathan.kurfali@su.se, dezeyrek@metu.edu.tr

Abstract

In this work, we introduce a lightweight discourse connective detection system. Employing gradient boosting trained on straightforward, low-complexity features, this proposed approach sidesteps the computational demands of the current approaches that rely on deep neural networks. Considering its simplicity, our approach achieves competitive results while offering significant gains in terms of time even on CPU. Furthermore, the stable performance across two unrelated languages suggests the robustness of our system in the multilingual scenario. The model is designed to support the annotation of discourse relations, particularly in scenarios with limited resources, while minimizing performance loss.

Keywords: Discourse Connectives, Gradient Boosting, linguistically-informed features

1. Introduction

Recent advancements in deep learning have significantly improved state-of-the-art performances in natural language processing (NLP), and discourse parsing is no exception. Yet, despite these performance gains, these models demand high computing resources, which greatly hinders their usability, as many researchers around the world still lack access. Moreover, these models often act as black-box solutions, without providing any linguistic/theoretical insights regarding the task at hand. In our current submission, we present a lightweight detection system for connectives, which are considered as one of the most important building blocks of discourse structure.

Among various approaches to discourse structure, such as RST (Mann and Thompson, 1987) and SDRT (Lascarides and Asher, 2007), PDTB (Prasad et al., 2014) remains the largest annotated dataset (Prasad et al., 2014) involving discourse-level annotations. PDTB adopts a connective-based approach, where connectives are the anchors of discourse relations that hold between two text spans that have an abstract object interpretation, such as propositions or eventualities (Prasad et al., 2014). The challenge lies in distinguishing between connectives that function as discourse connectives (DC) and those that do not, known as non-discourse connective (NDC) usage. Consider examples (1) and (2):

1. He went to Paris for a vacation and visited the famous Eiffel Tower.
2. He speaks English and French.
(from (Başbüyük and Zeyrek, 2023))

PDTB recognizes the *and* in the first example as a discourse connective whereas, in the second

example, it does not, as it simply links two noun phrases. Thus, the first step in the PDTB annotation process is the detection of the connectives with discourse usage in a given text piece. In the current work, we address this issue using a lightweight model that utilizes linguistic features to efficiently identify discourse connectives without the need for specialized hardware, such as GPUs, which are still not available to most researchers worldwide. We train and evaluate our model in two languages, English (PDTB 2.0) and Turkish (Turkish Discourse Bank (TDB) 1.0 (Zeyrek et al., 2013)). The contributions of our work are:

1. We introduce a fast machine-learning model that detects connectives.
2. We show that this model achieves results close to state-of-the-art models.
3. We argue that verb-based features are the most important aspects of our lightweight connective detection model.

The paper is structured as follows. In Section 2, we introduce two lines of research that deal with connective detection and briefly summarize recently developed discourse parsers that are shown to work in Turkish as well as English. Section 3 introduces our method, and Section 4 the experimental setting as well as the data and baselines. In Section 5 we evaluate our model, and finally, in Section 6 we draw some conclusions.

2. Related Work

Reflecting the overall trend in the field, the literature on discourse parsing can be roughly divided into two parts: the body of works before, and after

the emergence of neural networks (NNs). Before the solutions based on neural networks became the default approach, the methods relied more on traditional approaches such as feature engineering or annotation projection (Wellner and Pustejovsky, 2007; Pitler and Nenkova, 2009; Versley, 2010).

Following the deep learning revolution, led by the increase in the available computing power and the amount of data, NN-based solutions slowly replaced linguistic features, and more black-box approaches have become popular (Hooda and Kosseim, 2017; Kurfali, 2020; Kutlu et al., 2023). Most prominently, the recent DISRPT 2021 (Zeldes et al., 2021) and 2023 (Braud et al., 2023) shared tasks have received only transformer (Vaswani et al., 2017)-based solutions to a range of languages including English and Turkish (e.g., Gessler et al., 2021; Metheniti et al., 2023; Anuranjana, 2023), with the exception of the TMVM model by Dönicke (2021), which utilized linguistic features derived from syntactic trees. Gessler et al. (2021) also stands out by integrating linguistics features into transformers.

3. Approach

The proposed connective detection model takes raw natural language data as input and determines which tokens are connectives. The task is modeled as a three-way token classification task, where each token can belong to one of three categories:

- *O*: The token is not part of a connective span.
- *B-Conn*: The token marks the beginning of a connective span. It can represent the entire span of the connective, as in single-word examples like *because*, or the first word of a phrasal connective, such as *on* in *on the other hand*.
- *I-Conn*: The token is the second or a subsequent word in a phrasal connective, e.g., *other* in *on the other hand*.

A computationally cheap and fast explicit connective detection algorithm should use symbolic or traditional ML-based approaches instead of deep learning architectures. At the same time, the features used by ML-based algorithms should be produced by algorithms with a time complexity lower than the inference time complexity of the ML model. For this purpose, we preferred to use gradient boosting to train our model. Gradient boosting is an ensemble method determining the optimal predictive model to enable us to use the decision trees more effectively (Friedman, 2001).

This iterative algorithm starts with a naive prediction (mostly an average line) to capture the target

values. In the second iteration, the residual between this prior prediction and the observed targets is calculated and a decision branch is adapted to decrease the sum of residuals. Repeating this process until the sum of residuals is minimized gives us a final decision tree for our classification task. We use the XGBoost (Chen et al., 2015) library to implement gradient boosting on our datasets.

We decided to incorporate three groups of features to our model. The first group involves verb-based features. These are the main features for our model and involve:

- Whether any of the three words before and three words after a candidate token is verb or not.
- Whether the current word is verb or not.
- The token-based distance of the current word to the previous and the following verbs.

The second group of features involves word-based features consisting of features such as the capitalization of words, word length, and a unique ID assigned to each word in the data, all of which can be produced with $O(n)$ time complexity.

The last group of features includes position-based features, by which we could produce in $O(n)$ time complexity, too. These involve the position of the current word in the sentence, also including the length of sentences based on words.

We used the XGBoost library to train our model with gradient boosting. The XGBoost library offers a wide choice of parameters for gradient boosting. Thus, we performed parameter tuning on *learning_rate* (contributions of each tree to the final model), *max_depth* (maximum depth of each tree), *n_estimators* (number of trees generated by the model), *max_delta_step* (a parameter that is useful for imbalanced datasets by preventing the weights from updating too much) and *min_child_weights* (a parameter to control the overfitting problem) which we consider to be the most important ones among these parameters. We used the grid search algorithm (Chicco, 2017) to choose the most effective tuning among these three parameters. Grid search systematically runs the different combinations of parameters and uses cross-validation (Stone, 1974) to find the best combination based on the performance. Recognizing the limited size of our dataset, we applied 3-fold cross-validation in our experiments to ensure a balance between model training time and validation robustness.

The dataset suffers from severe imbalance as discourse connectives do not occur as often. To deal with this, we also train our models with the weighted loss. We used inverse frequency weighting to determine the label weights. That is, for each

| Model | Learning Rate | Max Depth | N Estimators | Max Delta Step | Min Child Weights |
|---------------------|---------------|-----------|--------------|----------------|-------------------|
| PDTB 2.0 | 0.2 | 8 | 500 | 4 | 1 |
| PDTB 2.0 (Weighted) | 0.30 | 8 | 400 | 4 | 1 |
| TDB 1.0 | 0.15 | 10 | 500 | 4 | 1 |
| TDB 1.0 (Weighted) | 0.15 | 8 | 400 | 4 | 1 |

Table 1: Best Parameters for PDTB 2.0 and TDB 1.0 Datasets. Weighted refers to the classifiers trained with the "weighted" loss.

| Dataset | B-Conn | I-Conn | O | Connective Proportion(%) |
|-------------|--------|--------|-----------|--------------------------|
| TDB | | | | |
| Training | 7,044 | 1,259 | 385,256 | 2.11 |
| Development | 773 | 130 | 45,939 | 1.93 |
| Test | 849 | 165 | 45,944 | 2.16 |
| PTDB | | | | |
| Training | 23,848 | 4,499 | 1,032,851 | 2.67 |
| Development | 953 | 159 | 38,656 | 2.80 |
| Test | 1,245 | 238 | 54164 | 2.67 |

Table 2: The distribution of labels in the datasets. Refer to Section 3 for the label definitions. The last column denotes the proportion of all connectives to the total number of tokens.

i in our dataset, we computed w_i as

$$w_i = \frac{N}{C \cdot n_i}$$

where N is the total number of instances, C is the number of unique classes and n_i is the number of instances belonging to class i .

Weighted loss is a method used in imbalanced data to ensure that minority class data points contributes more to the model. The idea behind weighted loss is to assign a higher weight to the minority class data points while assign a lower weight to the majority class data points when computing the loss. Thanks to this approach, mistakes on the minority class become more "costly" for the model, causing it to pay more attention to correctly classifying instances of the minority class.

The best parameters according to the grid search are provided in Table 1.

4. Experimental setting

4.1. Data

In our experiments, we followed the training, development, and test splits proposed in DISRPT 2021 (Zeldes et al., 2021) to facilitate direct comparison of our models with the state-of-the-art systems evaluated there. The Turkish data in DISRPT is sourced from TDB 1.0 (Zeyrek et al., 2013), while the English data is based on PDTB 2.0 (Prasad et al., 2008). The distribution of the labels in the

respective datasets are provided in Table 2. DISRPT data uses these datasets without any pruning. Thus, our models are trained to explicit discourse connectives including discontinuous connectives such as "if .. then", "either .. or", etc. in addition to continuous or single word connectives. Alternative Lexicalizations (AltLex) connectives are also included in these datasets. AltLexes are not connective on their own but can act as connective when combined as multi word expressions.

4.2. Baseline Models

To put our results into perspective, we compare our model's performance against the best-performing systems in DISRPT 2021 and 2023 shared tasks. Additionally, we report the performance of a vanilla BERT model fine-tuned on the training set¹ (Devlin et al., 2018), to represent the current go-to approach for performing this task. We follow the standard token classification procedure using the default parameters and report the average performance across four different runs. The BERT baseline also provides insights into the time efficiency of our model, as that information is not available for the other baselines. It should be noted that all baselines, except for TMVM, are based on deep neural networks.

¹We used the *bert-base-cased* for English and the BERTurk model (Schweter, 2020) for Turkish.

| Model | Precision (%) | Recall (%) | f-score (%) | Inference Time (sec) |
|--------------------------------------|---------------|------------|-------------|----------------------|
| DisCut2023 (Metheniti et al., 2023) | 95.49 | 91.89 | 93.66 | – |
| DiscoDisco (Gessler et al., 2021) | 92.93 | 91.15 | 92.02 | – |
| Segformers (Bakshi and Sharma, 2021) | 89.73 | 92.61 | 91.15 | – |
| DisCut (Ezzabady et al., 2021) | 93.32 | 88.67 | 90.94 | – |
| TMVM (Dönicke, 2021) | 85.98 | 65.54 | 74.38 | – |
| BERT Baseline | 92.63 | 91.88 | 92.25 | 3.13 |
| Our Model | 89.10 | 78.71 | 83.58 | 0.02 (1.33*) |
| Our Model (Weighted) | 70.00 | 86.02 | 77.19 | 0.02 (2.03*) |

Table 3: Comparison of the Baseline Models and Our Model over PDTB 2.0 Using DISRPT Data Splits. * denotes inference time on CPU for our lightweight model.

| Model | Precision (%) | Recall (%) | f-score (%) | Inference Time (sec) |
|--------------------------------------|---------------|------------|-------------|----------------------|
| DiscoDisco (Gessler et al., 2021) | 93.71 | 94.53 | 94.11 | – |
| DisCut2023 (Metheniti et al., 2023) | 92.34 | 93.21 | 92.77 | – |
| Segformers (Bakshi and Sharma, 2021) | 90.42 | 91.17 | 90.79 | – |
| DisCut (Ezzabady et al., 2021) | 90.55 | 86.93 | 88.70 | – |
| TMVM (Dönicke, 2021) | 80.00 | 24.14 | 37.10 | – |
| BERT Baseline | 92.36 | 92.89 | 92.62 | 5.09 |
| Our Model | 87.41 | 71.96 | 78.94 | 0.01 (1.17*) |
| Our Model (Weighted) | 82.42 | 82.33 | 82.38 | 0.01 (1.55*) |

Table 4: Comparison of the Baseline Models and Our Models over TDB 1.0 Using DISRPT Data Splits. * denotes inference time on CPU for our lightweight model.

5. Results and Discussion

5.1. Results

We evaluated the performance of our model using the official evaluation script of DISRPT 2021.² The evaluation criteria are based on exact span matching, meaning that partial detection of phrasal connectives, such as identifying "because" within "That's because", does not contribute to the overall accuracy. For each language, micro-averaged precision, recall, and F-scores are reported.

The results of our system for English and Turkish are provided in Table 3 and Table 4, respectively. Despite our model's simplicity and reduced complexity, it demonstrates competitive performance when compared against the strong baselines. The best performances achieved in English and Turkish are very close to each other, suggesting that the model is robust across languages with different linguistic characteristics. Moreover, it must be highlighted that our submission outperforms the feature-based baseline, TMVM, in both languages, with the difference in Turkish being almost three-fold. We believe that this finding demonstrates the effectiveness of our set of features and further justifies their applicability to different languages.

²<https://github.com/distrpt/sharedtask2021>

Switching to weighted loss led to mixed results. In Turkish, the weighted loss increased the overall performance by 3 points; however, in English, it had a negative effect. Yet, in both cases, weighted loss significantly increased the recall of our models as expected. These findings indicate that while the approach increases the model's ability to identify true positive cases, its impact on precision, hence the overall performance, is language-dependent and requires further investigation.

On the other hand, our models achieved inference speeds at least three times faster than the BERT Baseline, despite being run on a CPU, unlike the BERT model which was trained and evaluated on a GPU. When both models are run on a GPU, the difference becomes nearly 250 times. This confirms that our model is indeed computationally less demanding, making it suitable for scenarios with limited computational resources.

5.2. Feature Importance

After training our model, we performed a feature importance test to determine which features made the highest contribution to the detection of DCs in TDB 1.0 and PDTB 2.0. The most important features detected by our best models in two languages are listed in Figure 1, Figure 2.

| Connective | Number of Correct Predictions | | Number of Incorrect Predictions | | Accuracy (%) |
|----------------------|-------------------------------|--------------------|---------------------------------|---------------------|--------------|
| | True Positive (TP) | True Negative (TN) | False Positive (FP) | False Negative (FN) | |
| and | 204 | 619 | 21 | 40 | 93.10 |
| for | 11 | 403 | 1 | 10 | 97.41 |
| then | 11 | 2 | 2 | 3 | 72.22 |
| Once | 0 | 0 | 3 | 1 | 0 |
| ve (and) | 181 | 477 | 33 | 25 | 91.90 |
| için (for) | 90 | 88 | 20 | 2 | 89.00 |
| Sonra (After) | 15 | 2 | 4 | 2 | 73.92 |
| aksine (contrary to) | 0 | 1 | 0 | 2 | 33.33 |

Table 5: Error Statistics for Selected Connectives in English (above) and Turkish (below). The top two connectives are the most frequent ones; the bottom two are the most mispredicted that occur at least three times.

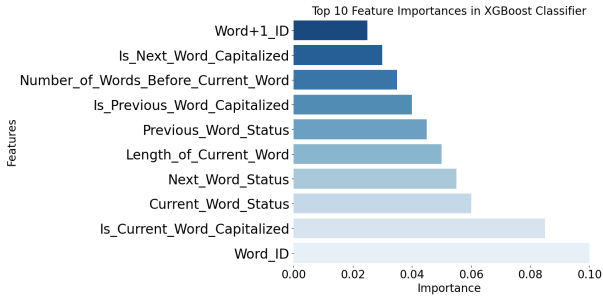


Figure 1: Feature importance in PDTB 2.0 for our best model

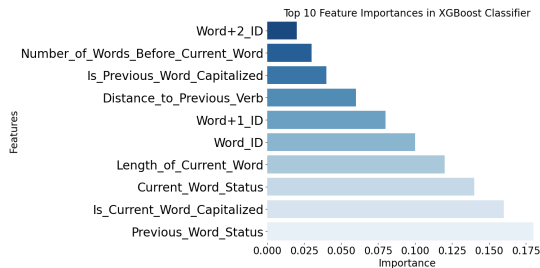


Figure 2: Feature importance in TDB 1.0 for our best model

As seen in the figures, word-based features such as Word ID and Capitalization check are prominent for PDTB. For TDB, the most critical feature is the information on whether the previous word is a verb (Previous_Word_Status). Additionally, while the status of the current word as a verb (Current_Word_Status) significantly contributes to the model for both languages, verb information of the next word for English and the previous word for Turkish stand out. We believe this may be attributed to the differences in word order between Turkish and English.

As shown in (Pitler and Nenkova, 2009), constituent tree-based features such as self category, parent category, sibling category provide very successful results in detecting explicit connectives. However, since annotated trees aligned with raw data are needed to derive these features, deriving

these features also has an additional annotation cost. In fact, since the annotation process of a dataset with the PDTB formalism is easier than the constituent tree annotation process, deriving the features to be used for automatic annotation may even cause higher costs than handmade annotation. This shows that our system, in addition to being lightweight compared to deep learning models, is also lightweight compared to classical approaches in terms of producing features effectively and at low cost.

5.3. Error Analysis

In this section, we discuss our model’s performance through error analysis. We present the error distribution for selected connectives in Table 5 and discuss some examples. The table highlights the first two connectives as those with the highest occurrence in our dataset, while the last two are identified as the most frequently mispredicted connectives above the specific threshold of 3. For Turkish data, the model tends to over-predict discourse connective (DC) usage over non-discourse connective (NDC) usage while in the PDTB, it is more cautious, often missing instances where connectives serve as DCs.

The examples below are provided to highlight the mistakes of our model. We show the mispredicted tokens by underlining, correctly predicted ones in bold fonts.

Example (4) showcases an unusual case where our model incorrectly identifies a noun in the Turkish dataset, *aklı* (‘mind’), as a discourse connective.

- Laiklik zaten, inançlara saygı duyarak aklı özgürleştirmektedir.(False Positive)
‘Secularism already means liberating the mind by respecting beliefs.’

This error is noteworthy because the sentence does indeed contain a connective that expresses a manner relation, specifically through the (intra-sentential) suffixal connective -arak attached to the verb preceding *aklı*. Yet, such suffixal connectives are later added to the TDB in its 1.2 version (Zeyrek

and Er, 2022) and are missing in the DISRPT training data. We have spotted several more cases exhibiting the same behavior which suggests that our model is generalizing to the connectives that are not seen in its training data.

Examples (5) and (6) illustrate one of the most common mistakes of our model, both in Turkish and English datasets. In Turkish, it includes a phrasal connective *zaman da* ('when' used with the focus particle); yet, our model only identifies the first part, *zaman* ('when'), missing the focus particle, *da*. In English, the system only recognizes *for*, missing the rest of the connective. Due to the strict evaluation strategy that requires an exact span match, this prediction is classified as misprediction. Overall, the phrasal connectives are particularly challenging.

5. Uygun düştüğü sanıldığı **zaman da** hemen birbirlerinin üzerinden kayıp gideceklerdi. (False Negative)
'When people thought [it] fits, they would immediately slip over each other'.
6. **For** instance, Gannett Co. posted an 11% gain in net income, as total ad pages dropped at USA Today, but advertising revenue rose because of a higher circulation rate base and increased rates. (False Negative)

6. Conclusion and Further Studies

In this study, we introduced a lightweight, gradient-boosting-based system for detecting discourse connectives, achieving competitive performance with significantly faster inference speeds compared to deep learning-based alternatives. Our approach demonstrated robustness across English and Turkish, indicating its utility in multilingual settings and scenarios with limited computational resources. Thanks to the speed and accuracy of our system, our model can be used to mine large amounts of data that can be used to facilitate the development of new discourse-annotated corpora or as the training data of discourse-focused language models.

7. Bibliographical References

Kaveri Anuranjana. 2023. Discoflan: Instruction fine-tuning and refined text generation for discourse relation label classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 22–28.

Sahil Bakshi and Dipti Misra Sharma. 2021. A transformer based approach towards identification of discourse unit segments and connectives.

In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 13–21.

- Kezban Başbüyük and Deniz Zeyrek. 2023. Usage disambiguation of turkish discourse connectives. *Language Resources and Evaluation*, 57(1):223–256.
- Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol T Rutherford, and Amir Zeldes. 2023. The disrpt 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21. ACL: Association for Computational Linguistics.
- Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.
- Davide Chicco. 2017. Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1):35.
- 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*.
- Tillmann Dönicke. 2021. Delexicalised multilingual discourse segmentation for disrpt 2021 and tense, mood, voice and modality tagging for 11 languages. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 33–45.
- Morteza Kamaladdini Ezzabady, Philippe Muller, and Chloé Braud. 2021. Multi-lingual discourse segmentation and connective identification: Melodi at disrpt2021. In *2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 22–32. ACL: Association for Computational Linguistics.
- Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. Discodisco at the disrpt2021 shared task: A system for discourse segmentation, classification, and connective detection. *arXiv preprint arXiv:2109.09777*.
- Sohail Hooda and Leila Kosseim. 2017. Argument labeling of explicit discourse relations using lstm

- neural networks. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 309–315.
- Murathan Kurfali. 2020. Labeling explicit discourse relations using pre-trained language models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 79–86. Springer.
- Ferhat Kutlu, Deniz Zeyrek, and Murathan Kurfali. 2023. Toward a shallow discourse parser for turkish. *Natural Language Engineering*, pages 1–26.
- Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.
- William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.
- Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. Discut and discret: Melodi at disrpt 2023. In *3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42. ACL: Association for Computational Linguistics.
- Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*.
- Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.
- Stefan Schweter. 2020. Berturk-bert models for turkish. *Zenodo*, 2020:3770924.
- Mervyn Stone. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83–82.
- Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments of discourse connectives. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 92–101.
- Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The disrpt 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12.
- Deniz Zeyrek, Işın Demirşahin, and Ayıışıǰı B Sevdik Çallı. 2013. Turkish discourse bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue & Discourse*, 4(2):174–184.
- Deniz Zeyrek and Mustafa Erolcan Er. 2022. A description of turkish discourse bank 1.2 and an examination of common dependencies in turkish discourse. *arXiv preprint arXiv:2207.05008*.

Shallow Discourse Parsing on Twitter Conversations

Berfin Aktaş, Burak Özmen

UFS Cognitive Science, University of Potsdam, Germany
berfinaktas@uni-potsdam.de, ozmen.brk@gmail.com

Abstract

We present our PDTB-style annotations on conversational Twitter data, which was initially annotated by Scheffler et al. (2019). We introduced 1,043 new annotations to the dataset, nearly doubling the number of previously annotated discourse relations. Subsequently, we applied a neural Shallow Discourse Parsing (SDP) model to the resulting corpus, improving its performance through retraining with in-domain data. The most substantial improvement was observed in the sense identification task (+19%). Our experiments with diverse training data combinations underline the potential benefits of exploring various data combinations in domain adaptation efforts for SDP. To the best of our knowledge, this is the first application of Shallow Discourse Parsing on Twitter data.

Keywords: shallow discourse parsing, Twitter, discourse relations, PDTB

1. Introduction

Discourse parsing, the identification of discourse relations between text spans, has seen substantial advancements in recent years. However, a significant challenge arises when the parsers are tested on a different domain, as recent research (Scholman et al., 2021; Liu and Zeldes, 2023) demonstrates a notable degradation in their performance. Consequently, the need for additional resources for discourse relations in diverse genres becomes increasingly important.

Penn Discourse Tree Bank (PDTB) refers to both the largest corpus, composed of news texts, annotated for shallow discourse relations and to the framework describing the annotation of these relations. The dataset (Prasad et al., 2018) is composed of written news texts (from Wall Street Journal). The main purpose of PDTB-style annotation is to identify two (mostly consecutive) arguments Arg1 and Arg2 which are semantically related. This relation can be constructed via explicitly expressed discourse connectives (i.e., an explicit relation) or can be inferred implicitly (i.e., an implicit relation).

There exist studies applying the PDTB framework to a variety of formal and informal spoken texts (Tonelli et al., 2010; Rehbein et al., 2016; Riccardi et al., 2016; Crible and Cuenca, 2017). These studies show that the use of discourse connectives and relations differs significantly between written and spoken data. Scheffler et al. (2019) conduct a pilot study on conversational Twitter data, where they annotated a corpus of Twitter Conversations (henceforth *TwiConv*) for explicit intra-tweet relations (i.e., the connective and arguments are in the same tweet). Their analysis indicates that Twitter conversations resemble spoken texts in terms of discourse relations. Nevertheless, there is still a noticeable gap in research focusing on interaction on

social media, which remains a relatively unexplored area.

Our primary contribution (Sections 2 and 3) is tackling this challenge through the expansion of the initial annotations put forth by Scheffler et al. (2019). In addition to existing explicit intra-tweet annotations (Example 1¹), we include in our annotations [A] explicit inter-tweet relations (i.e., arguments of the relation are located on different tweets, mostly posted by different users, as in Example 2), as well as [B] all implicit (Example 3) and [C] hypophora relations (i.e., question-answer pairs in the text as in Example 4).²

- (1) *Black folks in Alabama organized. And **WON!***
[Single Tweet]
- (2) Tweet1: *Like I said, you don't know the whole situation to make such a judgement.*
Tweet2: And **until you have raised one yourself, sit down and shut up!**
- (3) Tweet1: *[..] Time is short!!!*
Tweet2: **Not as short as your career highlights.** [..]
- (4) Tweet1: *Higher than a the office of a Governor?? Or he's talking of the offices when turned upside down?*
Tweet2: **A speaker is higher than the governor**

Our second contribution (Section 4) is the first, to the best of our knowledge, application of shallow discourse parsing on Twitter data. We apply

¹In the examples given in this paper, first argument (Arg1) in a discourse relation is marked by *italics letters*, second argument (Arg2) by **bold letters** and connectives by underlining.

²Annotations are available here: <https://github.com/berfingit/TwiConv-discourse-relations>

domain adaptation by retraining a state-of-the-art neural shallow discourse parsing model (Knaebel, 2021), using the annotations we generated.

2. Discourse Relations in TwiConv

2.1. Data

The TwiConv corpus contains English language tweets collected from the Twitter stream on several (non-adjacent) days in December 2017 and January 2018 without filtering for hashtags or topics. Conversations are gathered by recursively obtaining parent tweets, whose IDs were derived from the `in_reply_to_id` field of the tweet objects returned by the former Twitter API. For specifics regarding the data collection, refer to Aktaş and Kohnert (2020).

TwiConv comprises 1756 tweets, posted by 594 distinct users.³ Tweets are organized into 185 conversation threads⁴, with an average tweet length of 153 characters. The threads vary in length from 3 to 78 tweets, with an average length of 10 tweets and a median of 7. There are 48,172 tokens in TwiConv.

2.2. Annotation Procedure

Annotations were conducted by a linguistics undergraduate student. We built upon the guidelines devised by Scheffler et al. (2019), further extending them to encompass the additional relations we annotated. Additionally, we refined the instructions for selecting argument spans to enhance clarity for our annotators. Annotations were marked with the PDTB annotator tool (Lee et al., 2016). We followed the PDTB-3 scheme for annotations.

The PDTB-3 framework uses a 3-level hierarchy for the semantic categorization of relations (i.e., through sense labels), where at the top level is the “class” label, distinguishing between EXPANSION, COMPARISON, CONTINGENCY, and TEMPORAL relations. Level-2 and level-3 in the sense hierarchy represent the fine-grained labels refining the semantics of the class. There are a total of 36 categories available for assignment as sense labels. For more details on the PDTB sense hierarchy, see Webber et al. (2018).

2.3. Inter-annotator Agreement

We conducted an Inter-annotator Agreement (IAA) study on a subcorpus of 20 randomly chosen threads. They comprise 267 tweets with an average length of 187 characters. A second linguistics

³In the conversations, it is possible for a single user to respond multiple times.

⁴A set of tweets consisting of one or more users replying to each other is called a *thread* in our terminology.

student annotated them for the IAA computation. Following earlier PDTB studies (e.g., Prasad et al. (2008); Rehbein et al. (2016)), we report percent agreement for explicit relations on the sense assignments, Arg1 and Arg2 span selection, and for implicit relations on their senses.

The agreement on argument spans for **explicit** relations (Table 1) was notably high, surpassing those reported by Scheffler et al. (2019). This improvement is likely due to our less ambiguous span selection guidelines for social media symbols such as hashtags, links, and emoticons.

| Type | Exact | Partial |
|----------------------|-------|---------|
| Connective Detection | 71% | - |
| Arg1 Span | 79% | 93% |
| Arg2 Span | 95% | 97% |

Table 1: IAA for explicit relation text spans

Only the **implicit** relations annotated by both annotators were examined in this IAA study. We defined an implicit relation as shared between the two if both annotators identified an implicit relation with exactly matching argument spans. As a result, the argument spans (Arg1 and Arg2) for the implicit relations we analyzed always aligned. Therefore, our agreement analysis focused solely on the sense assignments for these shared implicit relations. Specifically, the first annotator identified 169 implicit relations, of which 126 shared argument spans with those identified by the second annotator. Hence, our agreement analysis is based on these 126 common implicit relations.

Table 2 presents the sense agreement statistics. IAA for implicit relations is generally lower compared to explicit relations, as found in existing literature (Prasad et al., 2008; Zeyrek and Kurfali, 2017; Zikánová et al., 2019; Hoek et al., 2021). Our statistics confirm the acknowledged difficulty in annotating implicit relations. Additionally, we argue that annotating implicit relations is particularly challenging in Twitter conversations due to the text ambiguity resulting from Twitter’s character limit (280 characters during data collection) and the non-standard items (e.g., hashtags, abbreviations, and images) in tweets.

| Sense Level | Explicit | Implicit |
|-------------|----------|----------|
| Level-1 | 88% | 68% |
| Level-2 | 82% | 45% |
| Level-3 | 76% | 41% |

Table 2: IAA for sense annotations

In Table 3 we present the most common disagreements in implicit relation senses between the annotators. Scholman et al. (2022) allow annotation of multiple senses and then determine the senses that

frequently occur together (p. 3287). We observe that the pair exhibiting the highest co-occurrence frequency in their study (*Conjunction* and *Result*) is identical to the one found in our disagreement matrix. Additionally, the pairing of *Arg2-as-detail* and *Conjunction* is another prevalent combination in both statistics. This suggests that our disagreements might correspond with the observations by [Scholman et al. \(2022\)](#), highlighting the inherent ambiguity of implicit relations and the necessity for implementing multi-sense annotation.

| Sense1 | Sense2 | Percentage |
|---------------|----------------|------------|
| Conjunction | Result | 9.7% |
| Belief.Reason | Reason | 6.9% |
| Conjunction | Arg2-as-detail | 5.6% |
| Contrast | Arg2-as-denier | 5.6% |
| Conjunction | Reason | 4.2% |
| Conjunction | Arg2-as-subst | 4.2% |

Table 3: Most common disagreements in sense assignments in implicit relation annotations

2.4. Quantitative Analysis

The annotations comprise a total of 2281 discourse relations, with 1237 originating from the prior annotations of [Scheffler et al. \(2019\)](#). Within the full set, 1433 are explicit relations, 732 are implicit relations, and the remaining 116 are hypophora relations.

We observe that explicit discourse relations are a frequent occurrence in our Twitter data. Out of 1756 tweets, 47% contain at least one discourse connective, and 22% contain more than one (up to 6). A tweet with 6 connectives is given in Example 5.

- (5) Yes, but if it were true and she has decided to run in 2020, it gives more people something to rally behind, a reason to get out and vote this year, a Democratic Congress when she arrives! I'm all in, and think an Oprah run would greatly help in 2018 Mid Terms!
#Oprah2020

Table 4 shows the distribution of intra- and inter-tweet relations. The majority of Explicit and Implicit relations occur within a single tweet, whereas Hypophora relations are typically inter-tweet relations. 98.5% of the inter-tweet relations span into two tweets, as illustrated in examples 3 and 4 for an implicit relation and an hypophora relation, respectively; but there also exist relation instances that span into three tweets (1.5%). Inter-tweet relations typically occur between tweets posted by different users (81%) but they also exist between tweets posted by the same user (19%).⁵

⁵A comparison of relations established by the same user and by different users is left to future work.

| Relation Type | intra-tweet | inter-tweet |
|---------------|-------------|-------------|
| Explicit | 90% | 10% |
| Implicit | 88% | 12% |
| Hypophora | 4% | 96% |

Table 4: Intra- and inter-tweet relation distributions (All relations except intra-tweet Explicit relations have been annotated by our team.)

3. TwiConv vs PDTB 3.0

Table 6 shows the distribution of the level-1 relation senses in our corpus and in the PDTB corpus ([Prasad et al., 2019](#)). Our Twitter data has substantially more CONTINGENCY relations than the PDTB. In line with this observation, connectives expressing CONTINGENCY relations like *if*, *when*, *because* and *so* occur relatively more frequently on Twitter as shown in Table 5. During our annotation process, we noticed that longer threads often represent argumentative discussions, and the prevalence of CONTINGENCY connectives can serve as evidence for this: Users provide substantiation for their arguments. In contrast, news texts in PDTB use more narrative (TEMPORAL) and EXPANSION relations.

| Connective | TwiConv | Connective | PDTB |
|------------|---------|------------|-------|
| and | 27.6% | and | 26.3% |
| but | 15.9% | but | 15.2% |
| if | 7.9% | also | 7.1% |
| so | 6.6% | if | 4.7% |
| when | 6.2% | when | 4.3% |
| because | 5.7% | while | 3.3% |
| or | 2.8% | as | 3.3% |
| also | 2.8% | because | 3.1% |
| as | 2.2% | after | 2.1% |
| then | 1.8% | however | 2% |

Table 5: Top ten connectives in the TwiConv and PDTB-3 explicit relations

Regarding the implicit/explicit difference, in the TwiConv corpus, CONTINGENCY relations are more often realized implicitly, whereas TEMPORAL relations are more often explicit (like in PDTB). In PDTB, COMPARISON relations are much more often explicit (25% vs 11%) whereas in the TwiConv data, both relation types have similar proportion.

Finally, we briefly look at patterns regarding spoken vs. written differences. [Crible and Cuenca \(2017\)](#) argue that discourse markers in spoken genres are more multi-functional than in written genres, which indicates greater diversity within spoken genres, particularly in the sense distributions of certain connectives. Here, we compared the

| Class | Relation | Twiconv | PDTB |
|-------------|----------|---------|------|
| EXPANSION | All | 32% | 44% |
| EXPANSION | Explicit | 33% | 42% |
| EXPANSION | Implicit | 30% | 46% |
| CONTINGENCY | All | 34% | 25% |
| CONTINGENCY | Explicit | 29% | 16% |
| CONTINGENCY | Implicit | 43% | 35% |
| COMPARISON | All | 24% | 18% |
| COMPARISON | Explicit | 25% | 25% |
| COMPARISON | Implicit | 23% | 11% |
| TEMPORAL | All | 10% | 13% |
| TEMPORAL | Explicit | 13% | 17% |
| TEMPORAL | Implicit | 4% | 8% |

Table 6: Level-1 sense distributions for TwiConv and PDTB 3.

level-1 sense annotations for “and” which is the most frequent connective in both corpora. Table 7 reveals that it is used to establish TEMPORAL relations (as illustrated in Example 6) in 8.2% of explicit relations in TwiConv, but is not used for that purpose in PDTB. Tonelli et al. (2010) had observed a similar pattern in their dialog annotation in Italian, where the connective “e” (“and”) can express TEMPORAL as well as EXPANSION relations. Furthermore, in TwiConv, the COMPARISON relations established by “and” are much more common than in PDTB (5.7% vs 0.03%). This supports the idea that TwiConv represents patterns of spoken language in terms of connective functionality, which we plan to study further in future work.

| Class | Twiconv | PDTB |
|-------------|---------|------|
| COMPARISON | 5.7% | 0.3% |
| CONTINGENCY | 4.0% | 2.7% |
| EXPANSION | 82.2% | 97% |
| TEMPORAL | 8.2% | - |

Table 7: Level-1 sense distributions for “and” (case insensitive)

- (6) [...] I’m going to create a totally new arbitrary number and assign meaning to it.

4. Shallow Discourse Parsing (SDP) on Twitter Conversations

Experiments. Our experiments utilize the neural shallow discourse parser “*discopy*”, which was introduced by Knaebel (2021). The *discopy* model achieves state-of-the-art results in connective identification, and also demonstrates competitive performance in other SDP tasks, notably in Arg1 identification. The experimental design was the one

proposed at the CoNLL Shared Task 2016 (Xue et al., 2016), and the reported results conform to that.

The main goal of our work is to assess whether incorporation of Twitter Conversation data into the training data of *discopy* affects the performance of the model when tested on TwiConv. To accomplish this, we segment our TwiConv data into training, testing, and validation sets with the distribution of 80%, 10%, and 10% of data, respectively.

We then combine the TwiConv training set with different portions of PDTB data from the CoNLL 2016 Shared Task (Xue et al., 2016), which consists of 930k tokens and has been employed to train the original *discopy* model. These combinations encompass varied token quantities from the PDTB data, allowing us to manipulate the proportion of TwiConv data in the training set. We establish four distinct setups:

- setup 1 (only PDTB)
- setup 2 (30k tokens PDTB + TwiConv)
- setup 3 (465k tokens PDTB + TwiConv)
- setup 4 (complete PDTB + TwiConv)

We conduct experiments in these setups with both RoBERTa- and BERT-base embeddings, and we show the results in Table 8. (We only present the best scores for the sake of simplicity.)

We also implemented preprocessing steps on TwiConv, which involve eliminating URLs, poster handles, mentions, and transforming hashtags into complete words. For instance, ‘#ClintonFoundation’ was changed to ‘Clinton Foundation’. The results for the same setups with the preprocessed data are also provided in Table 8.

Results. Our baseline consists of parsing our test set with the *discopy* model trained solely on PDTB data (i.e., setup 1). We achieved our best results with RoBERTa-base for that setting, so we have adopted it as our baseline. It shows a substantial drop when run on the Twitter data, losing almost 50% of the results reported by Knaebel (2021) for PDTB parsing.

We obtained the best results for most of the metrics with BERT-base with setup 4, which improves over the baseline in almost all cases, including a 6% increase in connective identification. With the preprocessed data, we obtained the best results in setup 4 for most of the metrics with RoBERTa-base.

Discussion. Incorporating Twitter data into the training set generally proves useful; however, there is no universal configuration that consistently outperforms the other setups across all metrics. In most cases, an increase in the volume of PDTB training data leads to metric enhancements, although exceptions exist. For instance, the most

| Setup | F1 _{conn} | F1 _{Arg1} | F1 _{Arg2} | F1 _{Sense} |
|-------------------------|--------------------|--------------------|--------------------|---------------------|
| Baseline-rb | 0.46 | 0.25 | 0.38 | 0.32 |
| setup 3-rb | 0.52 | 0.24 | 0.37 | 0.37 |
| setup 4-rb | 0.51 | 0.25 | 0.33 | 0.39 |
| setup 3-bb | 0.52 | 0.28 | 0.34 | 0.49 |
| setup 4-bb | 0.52 | 0.27 | 0.39 | 0.49 |
| setup 2-rb ^p | 0.52 | 0.17 | 0.29 | 0.33 |
| setup 4-rb ^p | 0.49 | 0.3 | 0.37 | 0.51 |
| setup 3-bb ^p | 0.51 | 0.29 | 0.37 | 0.41 |
| setup 4-bb ^p | 0.51 | 0.28 | 0.37 | 0.38 |

Table 8: Performance of *discopy* on the TwiConv test set, with RoBERTa-base (rb) and BERT-base (bb). We use strict measuring according to (Knaebel, 2021), i.e., a 0.9 threshold for overlap. The “p” superscript signifies experiments conducted on preprocessed data.

favorable result for connective identification on preprocessed data (0.52) emerges when TwiConv is integrated with a relatively small portion (30K) of PDTB data. This highlights the significance of experimenting with various data combinations in domain adaptation efforts, depending on the SDP subtask that is most relevant for a downstream purpose.

When evaluating the optimal outcomes, it is evident that connective (+6%) and Arg1 identification (+5%) shows notable improvements through retraining. Sense identification exhibits improvements across nearly all configurations compared to the baseline, with a remarkable (19%) improvement when the data is preprocessed. On the other hand, Arg2 identification shows minimal benefits and, in most cases, becomes worse, with the best scenario yielding only a modest (1%) improvement. The average improvement in preprocessed results is only marginally superior to the outcomes attained using BERT-base on non-preprocessed data.

5. Conclusions

We introduced non-explicit (implicit and hypophora) and inter-tweet explicit relations to the TwiConv corpus, which was initially annotated by Schefler et al. (2019) for intra-tweet explicit relations, almost doubling the amount of original annotations. Subsequently, we applied a neural Shallow Discourse Parsing model to the dataset, enhancing the model’s performance on TwiConv data through retraining. We conducted experiments utilizing both BERT and RoBERTa embeddings, and the best results were obtained using BERT on the unprocessed data. This resulted in improvements across all tasks, except for Arg2 identifica-

tion, which presents an interesting case requiring further investigation. Extensive preprocessing of the Twitter data results in only marginal improvements.

6. Acknowledgements

We thank the anonymous reviewers and Manfred Stede for their valuable observations and suggestions. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 317633480 – SFB 1287.

7. Bibliographical References

- Berfin Aktaş and Annalena Kohnert. 2020. *TwiConv: A coreference-annotated corpus of Twitter conversations*. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 47–54, Barcelona, Spain (online). Association for Computational Linguistics.
- Ludivine Crible and Maria Josep Cuenca. 2017. Discourse markers in speech: characteristics and challenges for corpus annotation. *Dialogue and Discourse*, 8(2):149–166.
- Jet Hoek, Merel C.J. Scholman, and Ted J.M. Sanders. 2021. Is there less agreement when the discourse is underspecified? In *Proceedings of the DiscAnn Workshop*.
- René Knaebel. 2021. *discopy: A neural system for shallow discourse parsing*. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.
- Alan Lee, Rashmi Prasad, Bonnie Webber, and Aravind K. Joshi. 2016. *Annotating discourse relations with the PDTB annotator*. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 121–125, Osaka, Japan. The COLING 2016 Organizing Committee.
- Yang Janet Liu and Amir Zeldes. 2023. *Why can’t discourse parsing generalize? a thorough investigation of the impact of data diversity*. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.
- R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The

- Penn Discourse Treebank 2.0. In *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.
- Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.
- Giuseppe Riccardi, Evgeny A. Stepanov, and Shammur Absar Chowdhury. 2016. [Discourse connective detection in spoken conversations](#). In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6095–6099.
- Tatjana Scheffler, Berfin Aktaş, Debopam Das, and Manfred Stede. 2019. [Annotating shallow discourse relations in Twitter conversations](#). In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 50–55, Minneapolis, MN. Association for Computational Linguistics.
- Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2021. [Comparison of methods for explicit discourse connective identification across various domains](#). In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 95–106, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.
- Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. [DiscoGeM: A crowd-sourced corpus of genre-mixed implicit discourse relations](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France. European Language Resources Association.
- Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. [Annotation of discourse relations for conversational spoken dialogs](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2018. The Penn Discourse Treebank 3.0 Annotation Manual. Report, The University of Pennsylvania.
- Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Atapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. [CoNLL 2016 shared task on multilingual shallow discourse parsing](#). In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.
- Deniz Zeyrek and Murathan Kurfalı. 2017. [TDB 1.1: Extensions on Turkish discourse bank](#). In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81, Valencia, Spain. Association for Computational Linguistics.
- Šárka Zikánová, Jiří Mírovský, and Pavlína Synková. 2019. Explicit and implicit discourse relations in the prague discourse treebank. In *Text, Speech, and Dialogue*, pages 236–248, Cham. Springer International Publishing.

8. Language Resource References

- Prasad et al. 2019. *Penn Discourse Treebank Version 3.0*. University of Pennsylvania. distributed via Linguistic Data Consortium: LDC2019T05, ISLRN [977-491-842-427-0](#).

Search tool for An Event-Type Ontology

Nataliia Petliak, Cristina F. Alcaina, Eva Fučíková, Jan Hajič, Zdeňka Urešová

ÚFAL MFF UK

Charles University, Faculty of Mathematics and Physics, Computer Science School

Malostranské nám. 25, Prague 1, Czech Republic

{petliak,alcaina,fucikova,hajic,uresova}@ufal.mff.cuni.cz

Abstract

This short demo description paper presents a new tool designed for searching an event-type ontology with rich information, demonstrated on the SynSemClass (Urešová et al., 2022) ontology resource (version 5.0, (Urešová et al., 2023)). The tool complements a web browser, created by the authors of the SynSemClass ontology previously. Due to the complexity of the resource, the search tool offers possibilities both for a linguistically-oriented researcher as well as teams working with the resource from a technical point of view, such as building role labeling tools, automatic annotation tools, etc.

Keywords: language resource, lexical semantics, ontology, event types, demonstration, search tools, user interface

1. Introduction

Attempts aiming at improving formalized knowledge representation have resulted in the development of a number of huge lexical databases. Some of them are being interconnected to facilitate cross-formalism or cross-language studies. An exemplary project of this kind is a well-known initiative connecting lexical semantic resources called SemLink (Stowe et al., 2021), which aims to link together different mostly verb-oriented lexical semantic resources via a set of mappings; specifically, mappings between different word senses and semantic roles, as well as annotated corpus data. SemLink uses an online presentation system called The Unified Verb Index (UVI)¹ which merges those links and web pages from five different NLP projects.

All ontologies mentioned in SemLink are in one way or another related to the SynSemClass ontology for which the search tool described in this article has been created and on which it has been tested. However, the character of the search engine described herein allows it to be applied to other event-type ontologies as well.

Conversely, the search tool has been inspired in part by the UVI index search and other tools as available e.g., in BabelNet (Navigli and Ponzetto, 2010, 2012).

The organization of this paper is as follows: Sect. 2 introduces the SynSemClass ontology. Sect. 3 describes the already existing web browser used for browsing the ontology. Sect. 4 describes the core of the paper, the search engine, its interface, and examples to demonstrate its capabilities. We conclude and draw future plans in Sect. 5.

¹<https://uvi.colorado.edu/>

2. The SynSemClass Ontology

For exemplifying the search tool, we have used the SynSemClass ontology.² There also exists a web-based browsing tool.³ SynSemClass can be downloaded in full as a set of XML files.⁴

We present here a short description of the resource, taken from (Urešová et al., 2020). SynSemClass is one of the lexical semantic oriented projects dealing with the most common form of expressing events and states, namely verbs, and semantic role labeling.

SynSemClass concentrates on the participants of these events or states and the relations between them. For these relations the term “semantic roles” is used. Unlike other resources representing verb semantics, such as PropBank(s) (Kingsbury and Palmer, 2002), WordNet(s) (Fellbaum, 1998), FrameNet(s) (Baker et al., 1998), or VerbNet (Bonial et al., 2012), SynSemClass has been designed from the start as an “inter-lingual” resource, representing cross-lingual meaning of verbs (currently) in English, Czech, German and Spanish, including links to 18 other external lexical-semantic resources in these languages. Furthermore, SSC maps the valency behavior of the class members to the set of semantic roles in each class (for more details see (Urešová et al., 2020)). The SynSemClass classes are meant to represent eventive concepts “universally,” i.e., to multiple, typologically diverse languages in a single resource.

SynSemClass in its current version 5.0 includes 1,546 classes with 15,790 Class Members (lexical units). All classes have Czech and English mem-

²<https://ufal.mff.cuni.cz/synsemclass>

³<https://lindat.cz/services/SynSemClass50/>

⁴<http://hdl.handle.net/11234/1-5230>

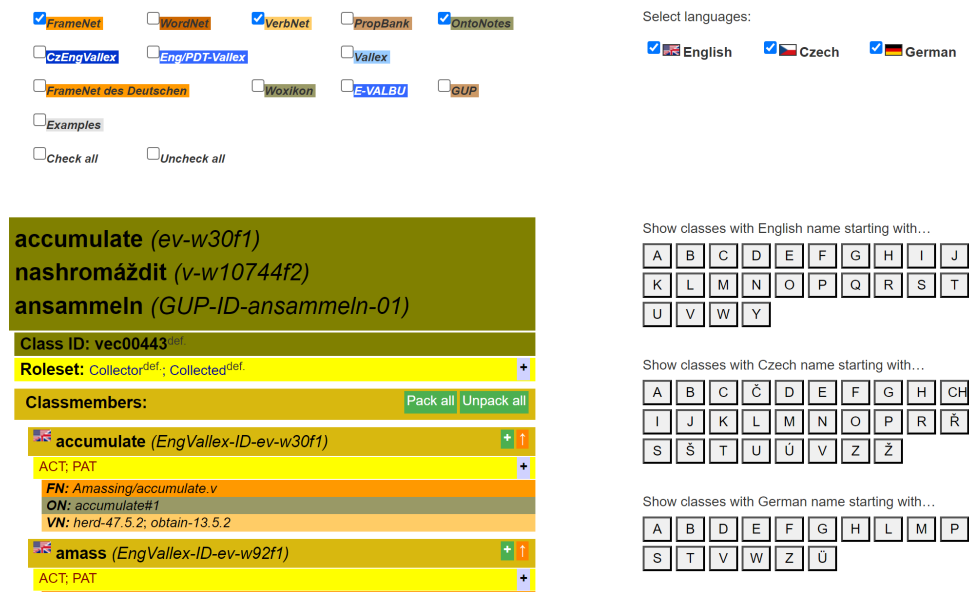


Figure 1: SSC web browser: (partial) display of results (*accumulate* class) as selected in the checkboxes

bers, and some of them also German and Spanish ones; the authors plan to extend it to more languages (Urešová et al., 2022).

3. The SynSemClass Web Browser

The existing web browser offers limited options for finding entries (it is truly just a browser), but it enables to selectively display all information contained in the ontology. Thus the search tool has been designed to make use of the browser for a fully formatted display of the search results based on user's post-filtering of the search results.

The web browser interface is divided into two parts: the contents are displayed on the left while the languages and entries to select from are on the right. The default view shows all contents, including external links, for all available languages. The user can select the resources and/or languages desired for browsing by ticking the box located in front of each resource label (Figure 1).

The entries to display are selected through the list(s) under "Show classes with [chosen language] name starting with..." The user first selects a letter and then a class by its name starting by that letter.

Additional information can be accessed by hovering over specific items. E.g., definitions of the class and the roles defined in the Roleset are displayed in a pop-up window when the mouse is over the superscript "def" located to the right of the Class ID and name of the role, respectively.

4. The Search Tool

4.1. The SSC Search Tool

The search tool⁵ offers multiple search criteria and flexible functionality for combining search options and building complex queries. The server-side development of the tool utilizes Express.js, which is retrieving data from the MongoDB database of the (converted) ontology data, instead of accessing the XML files directly. A React application is used on the client side.

Figure 2 depicts the tool's overall appearance.

4.2. Search Fields and Logic

The search interface contains the following search fields and query builders:

- **Lemma** - a class member's lemma, e.g., "bring".
- **Sense ID** - the sense ID of the class member, e.g., "EngVallex-ID-ev-w122f2".
- **Class ID** - common class ID to which the class members belong, e.g., "vec00107".
- **Filters** define the search languages. SynSemClass 5.0 currently includes English, Czech, German and Spanish, but the ontology will be expanded to include additional languages in the future. By default, all languages are searched; the user may select one or more languages to limit the search to those languages.

⁵<https://lindat.mff.cuni.cz/services/SynSemClassSearch/>; for the project description, API description and guidelines and general context see also <https://ufal.mff.cuni.cz/synsemclass/synsemclass-search-tool>

Filter languages English Czech German

Lemma search: Sense ID search: Class ID search:

Role(s) search:

Construct roles query in Conjunctive Normal Form,
 e.g., (Role1 OR Role2) AND (Role3 OR Role4) AND (Role5)

Select the role(s), drag and drop selected roles into the clauses brackets or between them.
 For mobile view, select the role; if needed, add more clauses and select more roles within each clause.

Figure 2: Search tool

Filter languages English Czech German

Lemma search: Sense ID search: Class ID search:

Role(s) search:

Construct roles query in Conjunctive Normal Form,
 e.g., (Role1 OR Role2) AND (Role3 OR Role4) AND (Role5)

Select the role(s), drag and drop selected roles into the clauses brackets or between them.
 For mobile view, select the role; if needed, add more clauses and select more roles within each clause.

Figure 3: Sample query: desktop view

- **The roles section** defines the search based on the roles of class members. The user can select a role from a list of available options. Conjunctive Normal Form (CNF) is used to perform a more advanced search of multiple role combinations. The user can manipulate the brackets (CNF clauses) by inserting roles between them and by adding additional clauses. The final query regarding roles is displayed at the bottom. It is always in CNF format, such as (Role1 OR Role2) AND (Role3). The desktop version of the web employs a drag-and-drop user interface for interactive and visual query creation. The mobile version of the application retains the capability to expand the roles query with additional drop-down options for each added clause.

Regular expressions can be used in any input field to match any strings. For example, “put.*” will show class members starting with “put”.

A sample query combining multiple search criteria is shown in Figures 3 and 4.

4.3. Presentation of Results

The summary of the results includes the number of class members and unique classes found in total, as well as by language (Figure 5).

The data representation is class-centric. Class members matching the criteria are grouped into their common classes, with essential information condensed at the top level.

The class ID is displayed at the top. The roles

are highlighted in green. Within a given class, only the top two class members with their respective sense IDs are displayed showing collapsed state of the result (Figure 6).

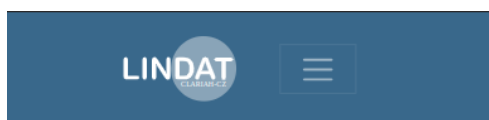
The “Show more” button reveals the complete list of class members with their sense ID, class member ID, and mapping (Figure 7).

Finally, expanding each individual class member with the arrow on the right reveals the member’s complete JSON content (Figure 8).

Both on the top-level presentation of results and inside fully expanded JSON content, the data contains links to external sources. For instance, clicking on a class ID or class name would highlight a link pointing to the corresponding class in the SynSemClass web browser. Similarly, clicking on the class member’s lemma would lead directly to a class member in the browser. In addition, in the fully expanded JSON output, highlighted fields also contain links to the corresponding entry in the external lexicon browsers (PDT-Vallex, EngVallex, etc.). For user convenience, class ID, sense ID and class member ID fields reveal a copy icon upon mouseover, allowing for a quick copy of the contents of these fields (Figure 9).

4.4. Search Tool API

The application was developed according to RESTful API specifications. By utilizing API endpoints, users can send search queries and receive rendered results identical to those obtained through the UI, or they can directly receive the raw response from the server for further processing



Filter languages English Czech German

Lemma search:

Sense ID search:

Class ID search:

Role(s) search:

Construct roles query in Conjunctive Normal Form,

e.g., (Role1 OR Role2) AND (Role3 OR Role4) AND (Role5)

Select the role(s), drag and drop selected roles into the clauses brackets or between them.

For mobile view, select the role; if needed, add more clauses and select more roles within each clause.

Objection

Objection x

(Agent x OR Recipient x) x

AND

(Offered x) x

(Agent OR Recipient) AND (Offered)

Figure 4: Sample query: mobile view

Found 9 class member(s) in 7 unique class(es).

- 9 English class member(s) in 7 class(es)
- 0 Czech class member(s) in 0 class(es)
- 0 German class member(s) in 0 class(es)

Figure 5: Search results information summary

class ID: vec00014

roles: Agent, Entity, State

class: keep (ev-w1792f2)

7 class member(s):

have (EngVallex-ID-ev-w1566f16_u_nobody)

hold (EngVallex-ID-ev-w1601f14_u_nobody)

...

Figure 6: Initial presentation of the result

class ID: vec00014

roles: Agent, Entity, State

class: keep (ev-w1792f2)

7 class member(s):

have (EngVallex-ID-ev-w1566f16_u_nobody) (vec00014-eng-cm00001) ACT ... Agent, PAT ... Entity, #alt[MANNLOC] ... State

hold (EngVallex-ID-ev-w1601f14_u_nobody) (vec00014-eng-cm00003) ACT ... Agent, PAT ... Entity, LOC ... State

keep (EngVallex-ID-ev-w1792f2) (vec00014-eng-cm00005) ACT ... Agent, PAT ... Entity, EFF ... State

keep (EngVallex-ID-ev-w1792f7) (vec00014-eng-cm00006) ACT ... Agent, PAT ... Entity, EFF ... State

keep (EngVallex-ID-ev-w1792f9) (vec00014-eng-cm00007) ACT ... Agent, PAT ... Entity, EFF ... State

hold (EngVallex-ID-ev-w1601f1) (vec00014-eng-cm00012) ACT ... Agent, PAT ... Entity, EFF ... State

keep (EngVallex-ID-ev-w1792f10) (vec00014-eng-cm00015) ACT ... Agent, PAT ... Entity, EFF ... State

Figure 7: Result showing all class members within the given language class

class ID: vec00382

roles: Communicator, Information, Audience, Addressee

class: emphasize (ev-w1124f1)

2 class member(s):

put (EngVallex-ID-ev-w2449f7_u_nobody) (vec00382-eng-cm00020) ACT ... Communicator, PAT ... Information, ... Audience, Addressee

```

@id: vec00382-eng-cm00020
@idref: EngVallex-ID-ev-w2449f7_u_nobody
@lang: eng
@status: yes
@lexidref: engvallex
@lemma: put
maparg:
  argpair:
    [0]:
      argfrom: @idref: vecargengACT
              form:
                spec:
    [1]:
      argfrom: @idref: vecroleCommunicator
              form:
                spec:
      argto: @idref: vecargengPAT
            form:
                spec:
      argto: @idref: vecroleInformation
  
```

Figure 8: Result with JSON content expanded

fine-tune (EngVallex-ID-ev-w1329f1)

(a) A copy icon for the Sense ID field

Copied EngVallex-ID-ev-w1329f1 to clipboard.

(b) Copied content

Figure 9: Quick copy tool

5. Conclusions and Future Work

We have presented a web-based search tool⁶ for searching an event-type ontology in general, demonstrated on the linguistically-motivated, richly cross-linked event-type ontology SynSemClass. This tool is aimed at researchers who want to explore the contents of the ontology, make comparisons across the linked external resources or across languages etc. It is linked to a web-based browser that shows all the contents of the ontology in detail, with further options to show or hide contents and to select languages, e.g. for comparison purposes, getting examples and graphical presentation of the results for research papers and other purposes. We will possibly also explore the currently available visualization tools for Linked Open Data, since the ontology could in principle be con-

⁶<https://lindat.mff.cuni.cz/services/SynSemClassSearch>

verted to them, but given the specialized nature and structured linking, it might not be possible without some loss of functionality.

The tool will be further developed and generalized to be able to configure it more easily to other similar resources or ontologies. Data will be added as the ontology grows. More complex queries will be designed and implemented, including one going across languages, for example. In the future, once annotated data becomes available, it will be integrated with corpus search.

6. Acknowledgements

The work described herein has been supported by the grant *Language Understanding: from Syntax to Discourse* of the Czech Science Foundation (Project No. GX20-16819X). It has been using data and tools provided by the *LINDAT/CLARIAH-CZ Research Infrastructure* (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

7. Bibliographical References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. *The Berkeley FrameNet Project*. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Claire Bonial, Weston Feely, Jena D Hwang, and Martha Palmer. 2012. Empirically Validating VerbNet Using SemLink. In *Seventh Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, Istanbul, Turkey.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA and London.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to PropBank. In *Proceedings of the LREC*, Canary Islands, Spain.

R. Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *ACL*.

R. Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

Kevin Stowe, Jenette Preciado, Kathryn Conger, Susan Windisch Brown, Ghazaleh Kazeminejad, James Gung, and Martha Palmer. 2021. *SemLink 2.0: Chasing lexical resources*. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 222–227, Groningen, The Netherlands (online). Association for Computational Linguistics.

Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2020. *SynSemClass linked lexicon: Mapping synonymy between languages*. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 10–19, Marseille, France. European Language Resources Association.

Zdeňka Urešová, Karolina Zaczynska, Peter Bourgonje, Eva Fučíková, Georg Rehm, and Jan Hajič. 2022. *Making a semantic event-type ontology multilingual*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1332–1343, Marseille, France. European Language Resources Association.

8. Language Resource References

Urešová, Zdeňka and Alcaina, Cristina Fernández and Bourgonje, Peter and Fučíková, Eva and Hajič, Jan and Hajičová, Eva and Rehm, Georg and Rysová, Kateřina and Zaczynska, Karolina. 2023. *SynSemClass 5.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, PID: <http://hdl.handle.net/11234/1-5230>.

Tiny But Mighty: A Crowdsourced Benchmark Dataset for Triple Extraction from Unstructured Text

Muhammad Salman^{1,*}, Armin Haller¹, Sergio J. Rodríguez Méndez¹, Usman Naseem²

¹School of Computing - CECC, The Australian National University, ACT, 2601, Australia

²School of Computing, Macquarie University, NSW, 2113, Australia

{Muhammad.Salman, Armin.Haller, Sergio.RodriguezMendez}@anu.edu.au, usman.naseem@mq.edu.au

* Corresponding Author

Abstract

In the context of Natural Language Processing (NLP) and Semantic Web applications, constructing Knowledge Graphs (KGs) from unstructured text plays a vital role. Several techniques have been developed for KG construction from text, but the lack of standardized datasets hinders the evaluation of triple extraction methods. The evaluation of existing KG construction approaches is based on structured data or manual investigations. To overcome this limitation, this work introduces a novel dataset specifically designed to evaluate KG construction techniques from unstructured text. Our dataset consists of a diverse collection of compound and complex sentences meticulously annotated by human annotators with potential triples (subject, predicate, object). The annotations underwent further scrutiny by expert ontologists to ensure accuracy and consistency. For evaluation purposes, the proposed F-measure criterion offers a robust approach to quantify the relatedness and assess the alignment between extracted triples and the ground-truth triples, providing a valuable tool for evaluating the performance of triple extraction systems. By providing a diverse collection of high-quality triples, our proposed benchmark dataset offers a comprehensive training and evaluation set for refining the performance of state-of-the-art language models on a triple extraction task. Furthermore, this dataset encompasses various KG-related tasks, such as named entity recognition, relation extraction, and entity linking.

Knowledge Graph (KG), Natural Language Processing (NLP), Text Annotation, Triple, Large Language Models (LLMs)

1. Introduction

Knowledge Graphs (KGs) have gained significant importance in a wide range of natural language processing (NLP) applications (Hogan et al., 2021). They serve as a valuable tool for organising information and extracting structured knowledge from unstructured data, such as plain text. Information in KGs is stored in a structured form, i.e., in the form of triples (subject, predicate, object), and the main source of information extraction is ‘text’, which is approximately 80% unstructured (Zong et al., 2021). Constructing KGs from unstructured text poses a challenge as KG requires the extraction of complete and accurate facts (triples) from the text. Many state-of-the-art KG construction methods have been developed, but they lack comparative analysis due to the unavailability of a benchmark dataset (Ji et al., 2021). To address this, a high-quality annotated dataset is essential for the evaluation of a model with competing techniques.

This paper introduces a novel dataset designed for triple extraction and validation from unstructured text. The dataset has been annotated by human annotators and verified by expert ontologists. It offers comprehensive coverage across various general domains and is enriched with high-quality annotations i.e., 96% verified. The dataset and evaluation criteria are publicly available¹ and can be leveraged

by the research community.

To construct our dataset, we used an open-source dataset (Zhang et al., 2020), which is mainly based on Wikipedia text and used for *Split and Rephrase* benchmarking. The general benchmark of ‘Small But Mighty’ serves our purpose because it contains compound and complex sentences and covers a wide range of textual domains. Before the annotation phase, we applied controlled sentence simplification techniques to the complex sentences contained within this dataset. This step ensured that the sentences were easily comprehensible for annotators, minimising the chances of missing any crucial fact during the labelling process. In the initial phase, a team of volunteer human annotators underwent training to label the text sentences according to our carefully developed annotation schema, which adhered to widely accepted standards in the field (Hogan, 2020). Subsequently, expert ontologists performed rigorous verification of the annotations in the second phase to maintain high quality and consistency. The workflow is shown in Figure-1.

Our dataset encompasses various KG construction tasks, such as entity recognition, relation extraction, and entity linking, all derived from unstructured text. We have created a valuable resource for researchers in NLP, information extraction, and related domains by employing a meticulous annotation process involving human annotators and expert

¹<https://w3id.org/salmon/TinyButMighty>

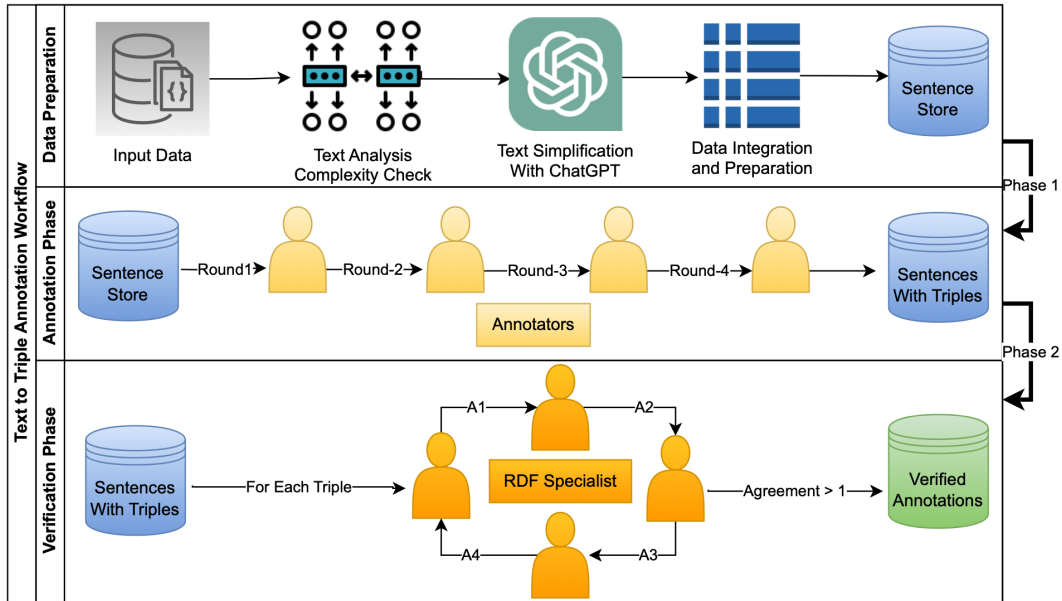


Figure 1: Workflow of Annotation Process

ontologists. The availability of this standardized dataset for KG construction from unstructured text, annotated and verified by experts, aims to stimulate further research and advancements in these critical areas of NLP. This paper has following **contributions**:

Web-based Annotation Tool: We have implemented a crowd-sourcing annotation system which can be used for multiple NLP annotation tasks.

Refinement of Existing Resource: We identified that the original simplified dataset still contains syntactic complexity and reduced that with OpenAI Generative Pre-trained Transformer (GPT), constructing a more robust simplified dataset.

Text-2-Triple Dataset: We were involved in a rigorous annotation process to create a novel dataset for triple extraction from unstructured text.

Evaluation Criterion: We have also proposed a triple-similarity evaluation criteria when the output triples are from unstructured text and cannot be identical to ground-truth triples.

2. Background and Related Work

Constructing KGs from unstructured text is a crucial task with wide-ranging applications in information retrieval, question answering, and other domains (Niklaus et al., 2018). KG construction from unstructured text has been a vibrant area of research in NLP (Zong et al., 2021; Heist et al., 2020; Gutiérrez and Sequeda, 2021), with various approaches proposed, including rule-based, machine learning-based, and hybrid methods that combine both (Hogan et al., 2021; Paulheim, 2017).

These approaches typically involve identifying entities (Delpeuch, 2019) and relations (Sakor et al., 2020) in the text and constructing a graph that represents the relationships between the identified entities (Wang and Yang, 2019). While several datasets have been utilised in KG construction approaches (Al-Moslmi et al., 2020), most of them were manually crafted for specific tasks or curated and selected from structured data sources such as Wikipedia or DBpedia (Kertkeidkachorn and Ichise, 2017; Liu et al., 2018). Such datasets pose challenges in training and evaluating KG construction models on unstructured text data, which is typically more diverse and noisy.

In Table 1, we reviewed triple extraction techniques and investigated the evaluation method. It shows that the developed techniques provide no state-of-the-art evaluation and rely on their own investigation. In the task of RDF triple extraction from structured or semi-structured text, the *WebNLG* (Gardent et al., 2017) dataset is being widely adopted for evaluation. To evaluate an RDF triple extractor, *WebNLG* can be used in reverse, i.e., instead of *RDF_Triple-Generated_Text*, *Generated_Text* can be used as input to identify the *RDF_Triples*. For relation extraction, GraphRel (Fu et al., 2019) applied the *WebNLG* and New York Times (NYT) (Riedel et al., 2010) datasets. The *NYT* dataset is generated from news articles and is also a widely adopted dataset for relation extraction (RE) tasks, but its limited relations (three entity types and 24 relations) restrict it from evaluating the KG construction system from unstructured text. Recently, REBEL (Cabot and Navigli, 2021) has been trained on four differ-

| Technique / Dataset | Target | Text Type | Evaluation |
|---|-------------------|----------------------|--------------|
| SEQ2RDF (Liu et al., 2018) | RDF Triple | Unstructured | \times |
| T2KG (Kertkeidkachorn and Ichise, 2017) | Triples | Natural Language | \times |
| FRED (Draicchio et al., 2013) | RDF, OWL | Natural Language | \times |
| Real Time RDF (Gerber et al., 2013) | RDF Triple | NEWS, Unstructured | \times |
| Exner System (Exner and Nugues, 2012) | RDF Triple | Wikipedia Articles | \times |
| UT2KB (Salim and Mustafa, 2021) | Relations | Unstructured | \times |
| Relation Extraction (Uddin et al., 2014) | Relations | Ebooks | \times |
| T2R (Hassanzadeh et al., 2013) | RDF Triple | Documents | \times |
| Seq2KG (Stewart and Liu, 2020) | RDF Triple | Domain Specific | \times |
| ER Extraction (Prasojo et al., 2016) | Entity, relations | Wikipedia Articles | \times |
| WebNLG (Gardent et al., 2017) | Text | RDF Triple | \times |
| NYT (Riedel et al., 2010) | RDF Triple | News | \times |
| GraphRel (Fu et al., 2019) | Relations | NEWS, Structured | \times |
| REBEL (Cabot and Navigli, 2021) | Relations | Semi-Structured | \times |
| CaRB (Bhardwaj et al., 2019) | Triple | Natural Language | \times |
| Our Benchmark | Triples | Unstructured Complex | \checkmark |

Table 1: State-of-the-Art methods For Triple Extraction and Evaluation Type

ent RE datasets and created a RE and classification dataset after fine-tuning and training on the BART-Large (Lewis et al., 2020) framework. CaRB (Bhardwaj et al., 2019) is also an addition to the community, but it seems limited in text domains and has not been qualitatively analysed.

Despite the success of information extraction approaches in different domains (Liu et al., 2023), there is still a need for high-quality annotated benchmarks for KG construction from unstructured text. This is particularly important as more complex and diverse data is becoming available in data lakes. We present a novel dataset for KG evaluation, providing a valuable resource for advancing the state-of-the-art in KG construction from unstructured text.

3. Dataset Creation and Annotation

The dataset creation and annotation workflow is shown in Figure-1, which involves the following steps.

3.1. Data Sources

To create our dataset, we started with an existing benchmark of complex sentences from IBM’s Split and Rephrase corpus (Zhang et al., 2020). This benchmark comprises over 720 compound and complex sentences from general text domains. We selected the general domain dataset for annotation so that it could best evaluate the models for unstructured text.

3.2. Data Refinement to Assist Phase-1 Annotators

The authors of the existing benchmark performed the “split and rephrase” function to transform complex sentences into simple sentences. However, our investigations noted that the existing simplification annotations are not robust enough for our purposes (shown in Table 2), as they often contain compound or complex sentences. After reviewing the potential limitation of the original corpus, we applied a method (Salman et al., 2023) to identify the syntactic complexity of the simplified text.

Based on our investigation, we developed a new “split and rewriting” module using GPT-3.5 (Floridi and Chiriatti, 2020), which enabled us to generate more accurate and meaningful simple sentences from the complex sentences in the dataset. In assessment, our pre-processed and re-annotated dataset has a more balanced distribution of complexity, as shown in Table 2. In this work, the sole purpose of sentence simplification is to assist Phase-1 annotators to get complete number of triple annotations. The sentence simplification is not part of Phase-2 and any further evaluation framework.

| Description | | Value |
|-----------------------|---------------|---------------|
| Complex Sentences | | 720 |
| Simplified Sentences | IBM Corpus | 3,565 |
| | GPT Annotated | 2,277 |
| Simplified / Sentence | IBM Corpus | 4.95 |
| | GPT Annotated | 3.16 |
| Performance | IBM Corpus | 90.15% |
| | GPT Annotated | 94.34% |

Table 2: Statistics of Complexity Measurement

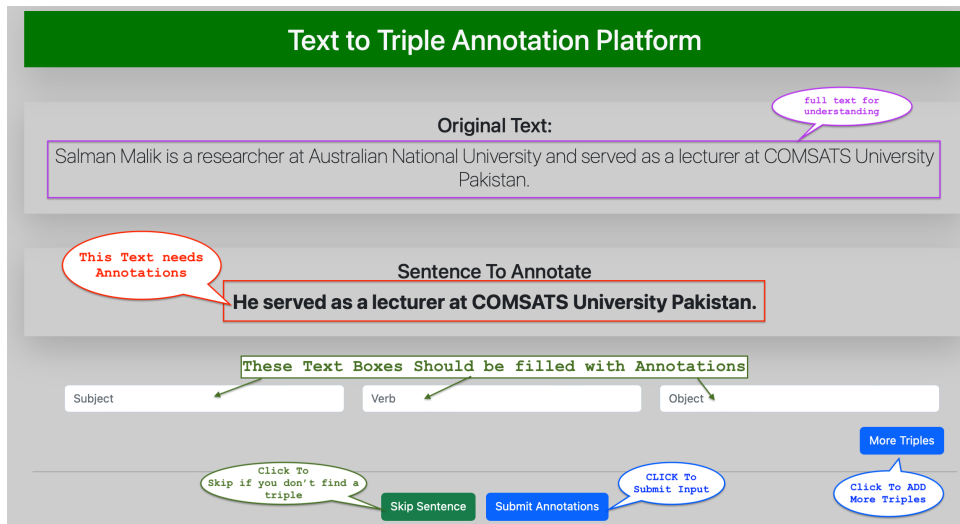


Figure 2: Platform Guidelines for Phase 1 Annotators

3.3. Annotation process: Phase 1

We invited participants through flyers in the department and ensured that no personal information was required in this process and that they could leave anytime. We recruited volunteer users to participate in the annotation process through a password-less Website, that does not require any personal information for sign-up. We enforced an *Exclusion Criteria* as well in which participants must have a command of the English language and the sentence structure of the language. We also described the annotation task to give the participants a brief understanding. These task-oriented briefings trained the participant for the annotation task. In the first phase of this annotation process, 127 participants voluntarily contributed to the task.

To ensure the quality and consistency of annotations, we followed a rigorous annotation process that involved four rounds of annotation by human volunteers. Each sentence went through four rounds of annotations, i.e., the round-1 annotations were evaluated by three participants in the following rounds. The annotators of each round were provided with clear guidelines and examples to ensure consistency in the annotation process and directed to a web-based system for annotations, as illustrated in the following section.

Web-based Annotation Tool For this task, we implemented a web-based tool to receive the annotations from participants. Ensuring the consistency of annotations, users are shown unique sentences in real-time with '*Concurrency Control*' while multi-user interaction with the platform. We recorded the elapsed time for each sentence's annotation. We deployed our web tool with Amazon Web Services (AWS) to ensure data safety and service reliability. The users were asked to ac-

cess the annotation tool through a website link and presented with a simple sentence and the original complex sentence. We asked users to mark/identify the maximum possible triples (subject, predicate, object) in each simple sentence.

During the start of each annotation session, participants were briefed about the annotation tool, and guidelines were supplied to get accurate annotations. On the website, we also provided them with a mock annotation exercise on how to use the tool's different features and provide annotations as shown in Figure-2. Following are some of the features of the website.

TEXT BOXES : Input for subject, verb, and object is taken separately in text boxes as shown in Figure-3

MORE TRIPLES : This button will add more text boxes if there is more than one triple in text.

SKIP SENTENCE : This button allows an annotator to skip a sentence for which there are no triples, as per the understanding of the annotator.

Annotation with Multiple Triples: A sentence is labelled with multiple triples contained in the sentence. Moreover, we encouraged the annotator to analyse '*Original Text*' to replace pronouns (he, she, and it, etc.) with the actual entity/noun label.

3.4. Challenges

During the annotation process, we analysed that participants were unfamiliar with KG concepts even though they were briefed, specifically about the notion of triples. To make it simple for the first round, we asked participants to identify the subject, verb, and object in the sample sentence. The resultant triples of the first annotation round were not of high quality and contained some "nonsensical facts". In the annotations, predicates were verbs only instead

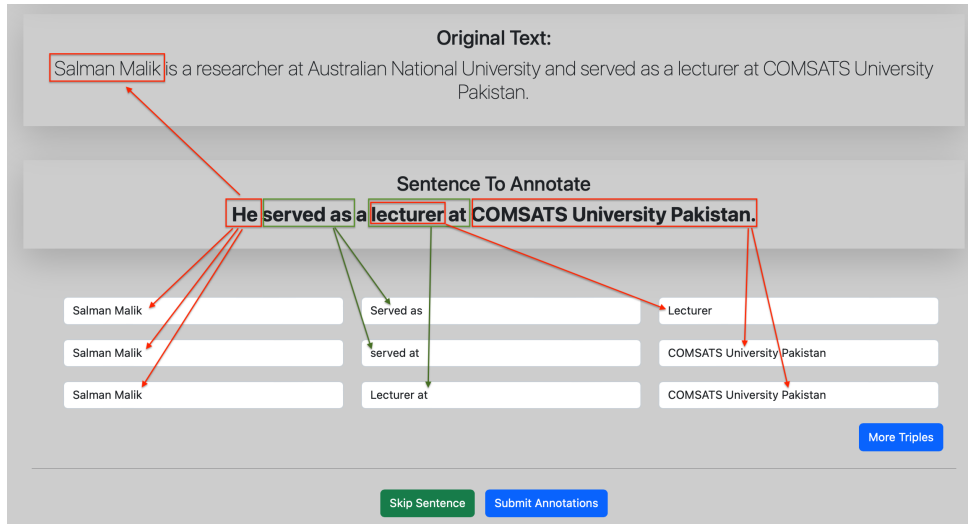


Figure 3: Entity Co-reference Resolution Handling

of proper prepositions, which sometimes drew confusion among the participants, e.g., *work* instead of *work in*, *work at*, *work from*, etc. Furthermore, we also identified that the participants were not incorporating the entity co-reference resolution in the annotations, i.e., the participants were reporting the pronouns (*He*, *She*, *It*, and *They*, etc.) from the simple sentence instead of referring to proper noun (name or tile) from the original text.

Therefore, we provided some further guidelines to get an improved version of annotations in the following rounds (the guideline for entity co-reference is shown in Figure-3). From the third round of annotation, we observed consistency in annotations between participants, which became more prominent in the fourth round.

3.5. Quality Assurance: Phase 2

To ensure the high quality of the annotations, we had each sentence annotated at least four times by different users in each round. We also provided clear guidelines and examples to the users to ensure consistency in the annotation process. After each round of annotation, we reviewed and refined the annotations to improve their quality and accuracy. We then finalised the annotations from the last round and performed a final entity co-reference resolution task to ensure consistency in the annotations across multiple sentences.

Finally, we involved expert ontologists in verifying each annotated triple's correctness and overall annotations based on the original sentence. This process ensured that the resulting dataset was accurate, reliable and suitable for training and evaluating KG construction models from unstructured text. For each annotated triple in Phase-1, we asked

the participants of Phase-2 to verify the quality of triples based on the following questions.

1. Is the annotated triple 'Correct'?
2. Is the annotated triple 'Partially Incorrect'?
3. Is the annotated triple nonsensical or vague?

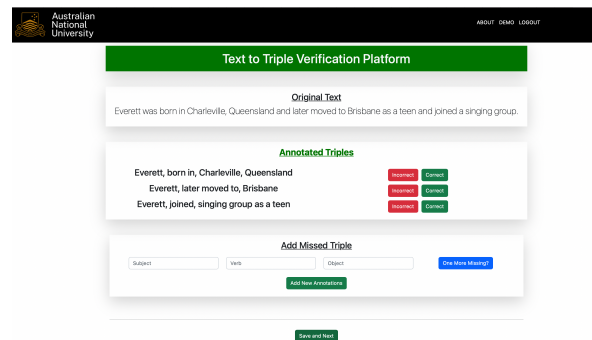


Figure 4: Annotation Verification Platform

4. Dataset Characteristics

The dataset comprises a collection of complex sentences, which are rewritten into more straightforward simplified sentences using OpenAI² as mentioned in Section 3.2. The original complex sentences were obtained from an already published benchmark dataset for text simplification (Zhang et al., 2020). The dataset comprises of 720 complex and 2,277 simple sentences, annotated with all possible triples (subject, predicate, object) by human volunteers, as explained previously.

²<https://openai.com>

| PHASE-1 | | PHASE-2 | |
|---|--------|---------------------------------------|---------------|
| # of Complex Sentences | 720 | Data Sample | 10% |
| # of Simplified Sentences | 2,277 | # of Triples in Data Sample | 317 |
| # of Annotators | 127 | # of Experts | 5 |
| # of Annotation Rounds | 4 | # of Correct Triples | 242 (76.34%) |
| # of Total Annotated Triples | 11,425 | # of Partially Correct Triples | 63 (19.87%) |
| # of Triples in Final Round | 3,736 | # of Nonsensical Triples | 12 (3.78%) |
| Avg. Triples Per Simplified Sentence | 1.64 | Verified Triples After Phase-2 | 305 |
| Avg. Triples Per Original Sentence | 5.18 | Success Rate of Annotations | 96.22% |

Table 3: Statistics of Annotated Dataset for Each Phase

Each simple sentence is part of the original complex sentence, and the annotations include all possible triples that could have been extracted from the simple sentence. The dataset covers a wide range of topics, including science, technology, history, and literature. The complexity of the sentences varies, ranging from moderately complex to compound and complex sentences, making it suitable for evaluating KG construction models from unstructured text.

4.1. Statistics

In this section, we provide the statistics of the annotation process and the contribution of volunteer participants. In phase-1, 127 volunteer annotators participated in the annotation process. We conducted four rounds of annotations; in each round, more than 2,200 annotations were recorded. A total of 11,425 triple annotation hits (add, edit, delete) were recorded. We observed refinement in the quality of annotations in each round and got 3,736 triple annotations for 720 complex sentences from the final round. The statistical summary of the dataset is presented in Table 3. We recorded an average of 5.18 triples per original sentence, depicting the complex nature of sentences. Moreover, each simplified sentence has an average of 1.64 triples, proving the meaning-preserving and fair dispersal of complexity after applying sentence simplification.

To verify the quality of phase-1 annotations, we invited expert ontologists to mark the sampled data as discussed in section 3.5. We removed the simplified sentence layer, and the participants were presented with the original sentence and its annotations only. We chose 80 unique complex sentences in our sample data that were labelled with 317 triple annotations in Phase-1. 76.34% of annotated triples are verified as 'Correct' while 19.87% were marked as 'Partially Incorrect' and edited with correct entity/predicate mentions. These updated triples were looped in the verification phase to take the agreement from other specialists and are marked as 'Correct'. Combining the verified triples, we have an agreement of verification on 96.22% by expert ontologists while

3.78% are marked as 'Incorrect or Nonsensical'. The relative distribution frequency of the correctness rating by expert ontologists is shown in Figure-5, and the statistics of Phase-2 sampled data are shown in Table 3.

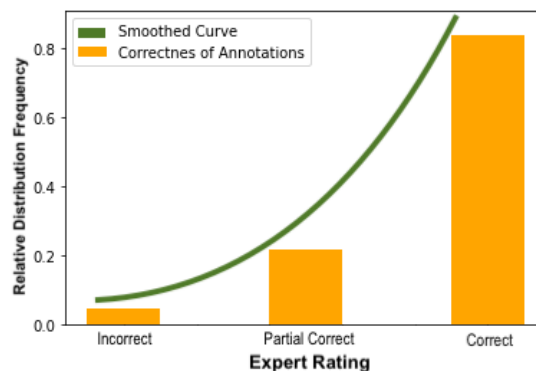


Figure 5: Relative Distribution of Correctness from Expert Ratings

4.2. Inter-Annotation Agreement (IAA)

The calculated Cohen's Kappa (κ) coefficients for Phase 1 revealed inter-rater reliability scores of 0.63, 0.78, and 0.84 for the rounds of annotation R1-R2, R2-R3, and R3-R4, respectively. The initial lower agreement in the R1-R2 round was attributed to inconsistencies observed in the use of the "is-a" relation and challenges in entity co-reference resolution, as analysed from the annotations in Phase 1 (Round #1). In response to these challenges, we introduced additional guidelines detailed in Section 3.4. The subsequent annotation rounds demonstrated improved consistency, evident from the enhanced κ scores observed between R2-R3 and R3-R4.

In Phase 2, the κ scores consistently exceeded 0.97, an authentication of the annotators' expertise within the research domain. The qualitative analysis phase agreed that an annotation must acquire at least two agreements to be deemed 'correct'. The overall score was thus established on annota-

Algorithm 1 F-Measure/Similarity of Triple-sets Calculation Criteria

```
1: Input: Triples  $T$  and Ground Truth  $GT$ 
2: Output: F-Measure
3: Coefficient:  $COSINE, JACCARD$ 
4: procedure CALCULATEFMEASURE( $T, GT$ )
5:   Initialize  $Relatedness(T)$  to 0
6:   Initialize  $Similarity(t_{ij})$  for each  $t_i$  in  $T$  and  $gt_j$  in  $GT$  to an empty list
7:   Initialize  $Adjustment$  as  $\max(LEN(T), LEN(GT))$ 
8:   for all  $t_i$  in  $T$  do
9:     Initialize  $Similarity_{t_i}$  as an empty list
10:    for all  $gt_j$  in  $GT$  do
11:       $Similarity(t_{ij}) \leftarrow Coefficient(t_i, gt_j)$ 
12:      Append  $Similarity(t_{ij})$  to  $Similarity_{t_i}$ 
13:    end for
14:     $Score_{t_i} \leftarrow \max(Similarity_{t_i})$ 
15:     $Relatedness(T) += Score(t_i)$ 
16:     $Relatedness(T) \leftarrow Penalty|Amnesty$  ▷ Based on Threshold
17:  end for
18:   $F\text{-Measure}(T) \leftarrow \frac{Relatedness(T)}{Adjustment}$ 
19:  return  $F\text{-Measure}$ 
20: end procedure
```

tions that attained consensus among the reviewers. This rigorous consensus requirement supports the reliability and validity of the annotation process, contributing to the exceptionally high agreement rates observed in Phase 2.

5. Experiments

5.1. Evaluation Criterion

Algorithm 1 calculates the F-Measure based on the given triples (T) and ground truth (GT) triples. T and GT sets are comprised of the model's output and verified annotations by experts, respectively. The algorithm iterates over each triple extracted by the model and calculates the similarity between each triple (t_i) and each ground truth (gt_j). The maximum similarity value is awarded as the score for a given t_i . The similarity of each triple is also awarded a penalty or amnesty based on thresholds. F-Measure is computed by taking the average of total similarity of T w.r.t. the maximum number of triples in T or GT to penalise the model for less/more generated triples. Finally, a penalty and amnesty are awarded again for final computation. In the penalising scheme, a triple is considered *Incorrect* if the relatedness is less than 50% of the ground truth triple. Conversely, in an amnesty scheme, an extracted triple is considered *Correct* if it has relatedness more than 80% with the ground truth triple.

In summary, our algorithm iterates over the triples and ground truths, computes similarity scores, accumulates the relatedness, and calculates the F-Measure. By incorporating penalties or amnesty

based on threshold values, the algorithm allows for flexible adjustment and evaluation of the relatedness and F-Measure.

$$SCORE(t) = \begin{cases} 1, & \text{if } Similarity_{t,gt} > 0.80 \\ Sim_{t,gt} & \text{otherwise} \\ 0, & \text{if } Similarity_{t,gt} < 0.50 \end{cases}$$

5.2. Baseline Methods

We applied some baseline techniques to our newly annotated dataset for preliminary evaluation. For this purpose, we selected two well-known language processing libraries and one generative pre-trained model (GPT-4).

SpaCy's SVO Model *Textacy* is designed on a high-performance and widely-used NLP library (SpaCy) for text processing. Specifically, we leverage *Textacy's* Subject-Verb-Object extraction feature for preliminary probes of our evaluation criterion and benchmark dataset.

CoreNLP OpenIE Triple Extractor Stanford's CoreNLP OpenIE triple extractor model was tuned to retrieve the maximum triples from a given sentence without any restrictions. While investigating the resulting triples, the Stanford OpenIE model generated non-informative triples for some sentences, such as [Airport, is, Located]. The implementation of Stanford OpenIE is publicly available³ and can be accessed and used in multiple ways, including as a Python library and wrapper.

³<https://github.com/philipperemy/stanford-openie-python>

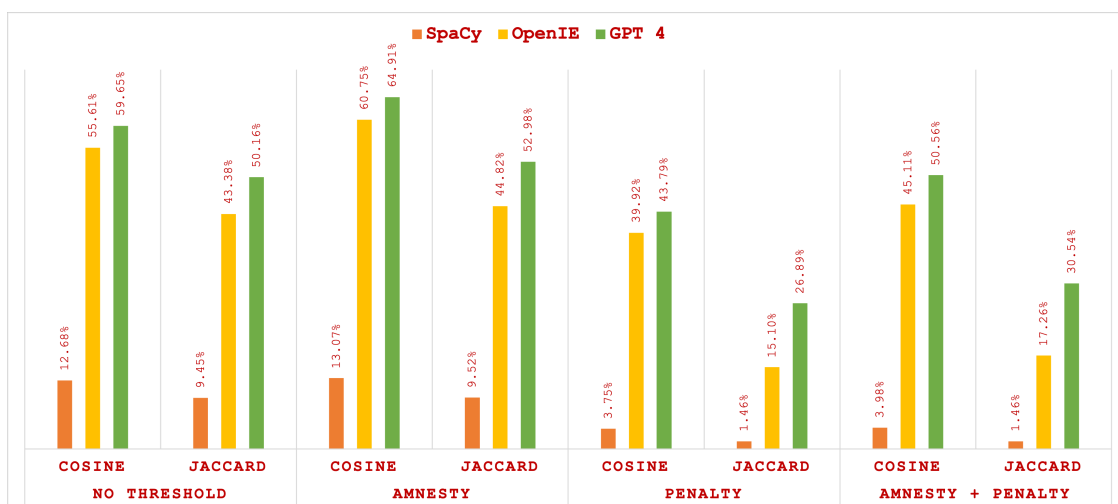


Figure 6: Performance of Models on Verified Sampled Data

GPT-4 Through OpenAI API For the purpose of evaluating triple extraction, we employed a GPT prompt-based approach. Specifically, we utilized the GPT-4 model as a triple extraction tool through *OpenAI* API that elicits relevant triples from the input text. After trying multiple query prompts, we settled on the best-resulting prompt for triples identification.

5.3. Results and Discussion

In this section, we will discuss the performance of our preliminary baselines w.r.t. evaluation framework illustrated in the prior section. As shown in Fig 6, GPT-4 is leading the baseline methods in both similarity coefficients. On the other hand, the performance of SpaCy’s SVO model is relatively low because of its incapability to deal with sentences of complex structure. CoreNLP OpenIE also performed very well and competed with the GPT-4 in all measures; however, we investigated that OpenIE generated 834 triples while the ground truth dataset contains 326 triples for sampled data. SpaCy and GPT-4 triple extractor models generated 59 and 261 triples, respectively. In our evaluation framework, we have taken care of the number of triples identified by the model and penalised a model’s output with the normalization of the overall relatedness score. We have also applied *Fuzzy Similarity* to evaluate the quality of extracted triples. In fuzzy ratio along with amnesty and penalty, SpaCy, OpenIE and GPT achieved 13.9%, 56.68%, and 60.81% respectively in triple qualitative analysis.

GPT-4 outperformed other baselines in all aspects of the evaluation. The quality of extracted triples from GPT-4 is also of high quality because it generated fewer (261 vs 326) triples than the ground truth data but still managed to lead the performance table in all measures. Although GPT is

the best-performing model but there is still a huge margin of improvement. The inclusion of this resource in the fine-tuning process can enhance the large language models’ (LLMs) understanding and generation capabilities, enabling them to generate more accurate and contextually appropriate triples in KG construction tasks. Furthermore, with its wide range of domain-specific and general knowledge triples, the dataset presents an opportunity to improve the accuracy, reliability, and contextual awareness of LLMs, ultimately benefiting a variety of downstream applications, including question-answering systems, information retrieval, and KG construction.

6. Conclusion

This work presents a novel dataset to evaluate the KG construction tasks from unstructured text. The dataset comprises a collection of compound and complex sentences, which have been annotated with possible triples (subject, verb, object) by human volunteers. Expert ontologists have verified the annotations to ensure their correctness and consistency. We have demonstrated the potential of this dataset by using it to evaluate KG construction models and tools. We also proposed an algorithm that offers a robust approach to quantifying the relatedness and assessing the alignment between extracted triples and ground truth data, providing a valuable tool for evaluating the performance of triple extraction systems. Similar to the relatedness score, we subjected the F-Measure to penalty and amnesty based on threshold values to account for the quality of the matches. The results show that our dataset can improve the performance of KG construction models, especially in terms of extracting complete and accurate triples from unstructured text.

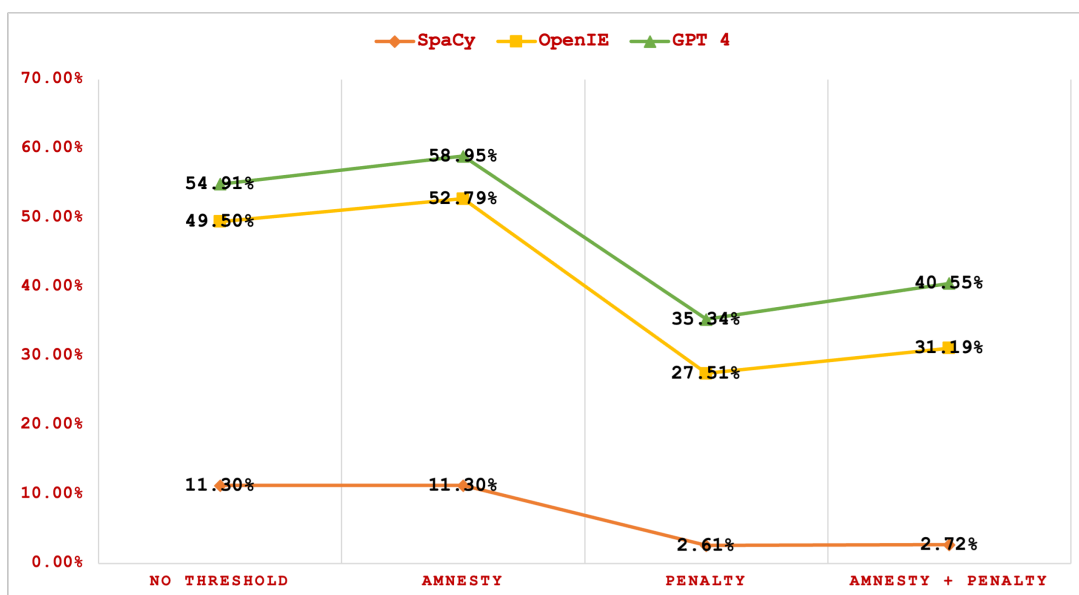


Figure 7: Average Performance of Models

The inclusion of this dataset in the fine-tuning process of an LLM can enhance its understanding and generation capabilities, enabling it to generate more accurate and contextually appropriate triples from unstructured text to construct a KG. Furthermore, the dataset facilitates the evaluation of triple extraction systems and contributes to advancing the research and development of NLP tasks related to knowledge graph construction and information extraction. With its wide range of domain-specific and general knowledge triples, the dataset presents an opportunity to improve the accuracy, reliability, and contextual awareness of LLMs, ultimately benefiting a variety of downstream applications, including question-answering systems, information retrieval, and knowledge graph generation.

Limitations

We acknowledge that there may be a need for larger datasets; however, as discussed in our evaluation of state-of-the-art models, the proposed benchmark dataset is sufficiently large to distinguish significant differences in accuracy for benchmark algorithms. The dataset is relatively small as compared to other text-related corpora and currently stands at 720 Complex and 2,277 simple sentences with high-quality annotations. However, this aspect may affect the dataset generalization, especially when training KG construction models that require a large amount of training data. Therefore, researchers should keep in mind the size of the dataset when using it and may need to supplement it with additional data if necessary. To address this limitation, this dataset will be used to fine-tune LLMs to make them capable of annotating large amounts of text.

This dataset is purely designed to evaluate the triple extraction system from unstructured text. However, it does not deal with RDF triples at this stage and we intend to transform the predicated as per RDF standards in our next release. We also intend to investigate the following research question while extending our dataset with Wikidata Mappings.

“Can KG construction from unstructured text data be improved by incorporating external knowledge sources such as a domain-specific ontology or open-domain knowledge bases?”

Ethics Statement

Since this study required human subjects to annotate and review the textual data (a "human-in-the-loop approach"), following a proper procedure to obtain ethical approval (protocol) was compulsory. A total of 127 participants were recruited to fulfil this purpose in Phase-1. Firstly, we obtained ethical approval for the annotation protocol from our University's research ethics committee. Under the approved research ethics (ANU Ethics Protocol 2022/464), we ensured the privacy and safety of participants. In other aspects of the protocol, volunteer participation to annotate the data is enforced, meaning there is no pressure or workload assignment from any course or program. Participants are also provided with the option to withdraw from the annotation process at any time. We also conveyed to the participants that the dataset would be publicly available to the research community.

References

- Tareq Al-Moslmi, Marc Gallofré Ocaña, Andreas L. Opdahl, and Csaba Veres. 2020. Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, 8:32862–32881.
- Sangnie Bhardwaj, Samarth Aggarwal, and Mausam Mausam. 2019. [CaRB: A crowdsourced benchmark for open IE](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6262–6267, Hong Kong, China. Association for Computational Linguistics.
- Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.
- Antonin Delpuch. 2019. Opentapioca: Lightweight entity linking for wikidata. *arXiv preprint arXiv:1904.09131*.
- Francesco Draicchio, Aldo Gangemi, Valentina Prezzutti, and Andrea Giovanni Nuzzolese. 2013. Fred: From natural language text to rdf and owl in one click. In *The Semantic Web: ESWC 2013 Satellite Events: ESWC 2013 Satellite Events, Montpellier, France, May 26-30, 2013, Revised Selected Papers 10*, pages 263–267. Springer.
- Peter Exner and Pierre Nugues. 2012. Entity extraction: From unstructured text to dbpedia rdf triples. In *WoLE@ ISWC*, pages 58–69.
- Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.
- Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1409–1418.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planning. In *55th annual meeting of the Association for Computational Linguistics (ACL)*.
- Daniel Gerber, Sebastian Hellmann, Lorenz Bühmann, Tommaso Soru, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2013. Real-time rdf extraction from unstructured data streams. In *The Semantic Web—ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I 12*, pages 135–150. Springer.
- Claudio Gutiérrez and Juan F Sequeda. 2021. Knowledge graphs. *Communications of the ACM*, 64(3):96–104.
- Kimia Hassanzadeh, Marek Reformat, Witold Pedrycz, Iqbal Jamal, and John Berezowski. 2013. T2r: System for converting textual documents into rdf triples. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 3, pages 221–228. IEEE.
- Nicolas Heist, Sven Hertling, Daniel Ringler, and Heiko Paulheim. 2020. Knowledge graphs on the web—an overview.
- Aidan Hogan. 2020. Resource description framework. *The Web of Data*, pages 59–109.
- Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37.
- Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*.
- Natthawut Kertkeidkachorn and Ryutaro Ichise. 2017. T2kg: An end-to-end system for creating knowledge graph from unstructured text. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

- Yue Liu, Tongtao Zhang, Zhicheng Liang, Heng Ji, and Deborah L McGuinness. 2018. Seq2rdf: An end-to-end application for deriving triples from natural language text. In *CEUR Workshop Proceedings*, volume 2180. CEUR-WS.
- Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. *arXiv preprint arXiv:1806.05599*.
- Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.
- Radityo Eko Prasajo et al. 2016. Entity-relationship extraction from wikipedia unstructured text. In *DC@ ISWC*, pages 74–81.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.
- Ahmad Sakor, Kuldeep Singh, Anery Patel, and Maria-Esther Vidal. 2020. Falcon 2.0: An entity and relation linking tool over wikidata. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3141–3148.
- Mustafa Nabeel Salim and Ban Shareef Mustafa. 2021. Uttokb: a model for semantic relation extraction from unstructured text. In *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 591–595. IEEE.
- Muhammad Salman, Armin Haller, and Sergio J Rodríguez Méndez. 2023. Syntactic complexity identification, measurement, and reduction through controlled syntactic simplification. *arXiv preprint arXiv:2304.07774*.
- Michael Stewart and Wei Liu. 2020. Seq2kg: an end-to-end neural model for domain agnostic knowledge graph (not text graph) construction from text. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 17, pages 748–757.
- Ashraf Uddin, Rajesh Piriyani, and Vivek Kumar Singh. 2014. Information and relation extraction for semantic annotation of ebook texts. In *Recent Advances in Intelligent Informatics: Proceedings of the Second International Symposium on Intelligent Informatics (ISI'13), August 23-24 2013, Mysore, India*, pages 215–226. Springer.
- XiuQing Wang and ShunKun Yang. 2019. A tutorial and survey on fault knowledge graph. *Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health*, pages 256–271.
- Li Zhang, Huaiyu Zhu, Siddhartha Brahma, and Yunyao Li. 2020. **Small but mighty: New benchmarks for split and rephrase**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1198–1205. Association for Computational Linguistics.
- Chengqing Zong, Rui Xia, and Jiajun Zhang. 2021. Information extraction. In *Text Data Mining*, pages 227–283. Springer.

Less is Enough: Less-Resourced Multilingual AMR Parsing

Bram Vanroy, Tim Van de Cruys

KU Leuven

Oude Markt 13, Leuven, Belgium

bram.vanroy@kuleuven.be, tim.vandecruys@kuleuven.be

Abstract

This paper investigates the efficacy of multilingual models for the task of text-to-AMR parsing, focusing on English, Spanish, and Dutch. We train and evaluate models under various configurations, including monolingual and multilingual settings, both in full and reduced data scenarios. Our empirical results reveal that while monolingual models exhibit superior performance, multilingual models are competitive across all languages, offering a more resource-efficient alternative for training and deployment. Crucially, our findings demonstrate that AMR parsing benefits from transfer learning across languages even when having access to significantly smaller datasets. As a tangible contribution, we provide text-to-AMR parsing models for the aforementioned languages as well as multilingual variants, and make available the large corpora of translated data for Dutch, Spanish (and Irish) that we used for training them in order to foster AMR research in non-English languages. Additionally, we open-source the training code and offer an interactive interface for parsing AMR graphs from text.

Keywords: AMR parsing, abstract meaning representation, semantics, corpora

1. Introduction

Abstract Meaning Representation (AMR, Section 2; [Banarescu et al., 2013](#)) is a meta-language for describing the semantic content of natural language sentences. It is agnostic to surface form (syntactic and lexical) and attempts to capture the meaning of a sentence in its most abstract form. While nodes are technically labelled with a linguistic form (typically a lemma optionally with a sense ID), these may as well be represented as an arbitrary identifier because they refer to a “meaning” rather than a lexical realisation of a meaning. Thanks to its machine-readable data format (as a directed, rooted graph, or as a sequence of triples) AMR has been employed for a variety of natural language processing (NLP) purposes (Section 3). However, the application of AMR to languages other than English has been stymied by the scarcity of large, annotated datasets that are suitable in size for training deep learning systems. AMR corpora exist, notably the English AMR 3.0 corpus ([Knight et al., 2020](#)), but manual annotation is costly and time-consuming. This means that AMR data sources are scarce, particularly for non-English languages.

The issue of resource scarcity is not only confined to languages that are commonly considered low-resource. Even languages like Dutch, which enjoys a relatively higher degree of digital presence and is spoken by around 24 million people, face challenges in annotated data for specialised tasks such as AMR. Even for Spanish, the fourth most spoken language in the world, there is a lack of suitable datasets for building deep learning systems for this task. In terms of task-specific re-

sources, such languages are still less-resourced – their mid-to-high resource nature in the traditional sense unfortunately does not transfer to a high availability of annotated data for all NLP tasks. Addressing this scarcity in terms of data availability, models, and research is crucial for the democratisation of NLP technologies and to ensure that the benefits of automating semantic AMR parsing is not confined to English.

In this context, to seek alternative approaches for performing, non-English text-to-AMR systems, multilingual models offer a promising avenue for exploration. Not only are these models computationally more efficient (training one multilingual model is more economical than training multiple monolingual ones); they also offer the advantage of easier deployment, as a single model can handle multiple languages. This efficiency is particularly salient in scenarios where computational resources are limited, a common situation in academic research and in deployments in less-resource environments. Moreover, multilingual models can be less data-hungry when training for each individual language, thereby partially mitigating the issue of data scarcity, which is the main topic of this paper.

This paper aims to investigate the efficacy of multilingual models in the task of text-to-AMR parsing, focusing particularly on English, Dutch, and Spanish. English serves as a well-resourced Germanic language, boasting a large, human-annotated AMR corpus (around 60,000 entries; [Knight et al., 2020](#)). In contrast, Dutch (also Germanic) and Spanish (Romance) are “less-resourced languages” in terms of AMR resources.

While both languages are widely spoken, they lack annotated and sizeable AMR corpora suitable for machine learning. However, their otherwise higher-resource status does allow for high-quality, automated machine translation (MT). We therefore make use of state-of-the-art MT systems to generate silver datasets for these languages, which we can then use to train deep learning systems (Section 4). We make available these datasets for other researchers as a tangible contribution.

We empirically and statistically evaluate multilingual (English, Spanish, Dutch) models under various configurations and compare them with monolingual counterparts to understand the trade-offs involved in terms of performance on the one hand and computational and data efficiency on the other. Specifically, we gauge how large the performance gap is between monolingual, full-resource models compared to artificially limited-resource, multilingual ones that have been trained on a subset of the data, and other multilingual models that were trained on the combined, full datasets of all languages. Our objective therefore is not to set new state-of-the-art results, although to the best of our knowledge our Dutch models are the best single-model text-to-AMR parsers for Dutch. Instead we offer insights into the advantages and disadvantages of multilingual text-to-AMR parsing and scrutinise the impact of data scarcity.

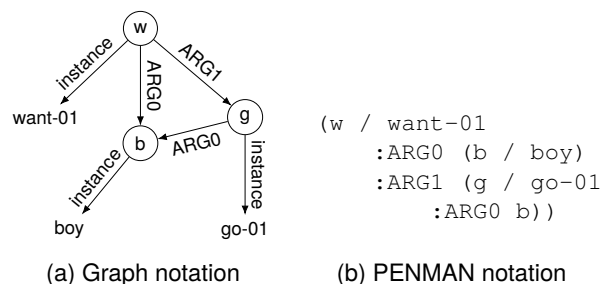
We provide valuable resources for the broader research community by publishing the models (monolingual models for English, Spanish and Dutch, as well as multilingual ones), the translated datasets for Dutch, Spanish, and Irish Gaelic (the latter not used in this paper but mentioned because it is part of our data release), the training and processing code, and an online interface to generate graphs from text.¹

2. Abstract meaning representation

AMR describes the meaning of a sentence in terms of “who does what to whom”, in an abstract form that is not bound by lexical or syntactic overt realisations. Therefore different sentences with the same meaning should have the same AMR realisation. AMR can be written as a directed, rooted graph (Figure 1a), e.g. the meaning of a sentence such as “The boy wants to go.” can be denoted with variables that can be used for (co)reference, such as w , b and g . Leaves in the graph are *concepts* so that the variable g refers to the concept $go-01$. These concepts are English words, special entities, or PropBank frame-sets (Kingsbury and Palmer, 2002), identifiable by

¹All resources can be found here: <https://github.com/BramVanroy/multilingual-text-to-amr>

their sense identifiers, such as $want-01$, which refers to the first meaning of *want* in the PropBank.² Special entities that are specific to AMR include concepts such as $phone-number-entity$ and $world-region$. For an exhaustive description of AMR, see the annotation guidelines.³



(<P1> want-01 :ARG0 (<P2> boy) :ARG1 (<P3> go-01 :ARG0 <P2>))

(c) Depth-first linearisation following Bevilacqua et al. (2021) (cf. Section 4.1)

Figure 1: AMR notations for the sentence “The boy wants to go.”. Adapted from Banarescu et al. (2013)

The edges in an AMR graph are labelled with the relationships between two nodes, or, rather, the role of the targeted node. Such relationships can be frame arguments that follow PropBank (such as the ARG_n roles); general semantic roles such as $:condition$ or $:accompanier$; quantities such as $:quant$ or $:unit$; date entities like $:day$ or $:decade$; and enumerations of different operators in $:op$ roles.

An AMR graph can be considered as logical triples of the following types of information: relationships, variables and concepts. Each triple is of the type $role(source, target)$ (e.g. $instance(w, want-01)$ or $:ARG0(w, b)$).

While the graph notation (and the underlying logical triples) is intended for computational readability, AMR can also be written in PENMAN notation (Matthiessen and Bateman, 1991), which makes it easier to read and write (Figure 1b).

3. Related research

3.1. Datasets

The English-oriented AMR 2.0 and 3.0 corpora (Knight et al., 2017, 2020) have been the cornerstone of much progress in English AMR generation and parsing. These datasets have been made

²<https://github.com/propbank/propbank-frames/tree/main/frames>

³<https://github.com/amrisi/amr-guidelines/blob/master/amr.md>

available through the Linguistic Data Consortium.⁴ AMR 2.0 contains 39,260 AMR annotations within the domain of news and weblog data. AMR 3.0 expands on that with 59,255 annotations in total, containing broadcasts and weblogs but also literary translations and Wikipedia articles. For multilingual purposes, the test set of the AMR 2.0 corpus has been partially translated to Spanish, German, Italian and Chinese Mandarin (1371 sentences per language; [Damonte and Cohen, 2020](#)), specifically for cross-lingual parsing. In this corpus, descriptively called “AMR 2.0 – Four Translations”, only the English source sentences were translated – the AMR structures remained unchanged. While such resource has been proven useful in multilingual research on AMR, its small size prohibits larger-scale experimentation and applicable.

In this work, we are interested in generating AMR for English but also for Dutch and Spanish. To the best of our knowledge, manually created or verified AMR corpora do not exist for Dutch. For Spanish, in addition to the limited translated AMR 2.0 partition mentioned above, laudable, manual efforts exist to create language-specific corpora. For instance, [Migueles-Abraira \(2017\)](#) annotated 50 sentences from Antoine de Saint-Exupéry’s novella *The Little Prince* translated into Spanish. [Wein et al. \(2022\)](#), on the other hand, defined annotation guidelines for Spanish and applied those guidelines to 486 Spanish sentences from the aforementioned “Four Translations” corpus to create a small but manually annotated gold corpus of Spanish AMR.

To collect multilingual data for AMR-to-text generation, [Fan and Gardent \(2020\)](#) were inspired by the methodology of [Damonte and Cohen \(2018\)](#) to make use of Europarl to create synthetic multilingual data. Europarl is very domain-specific and contains sentence-aligned parliamentary debates for English and many EU languages. The authors first automatically generate AMR from English sentences in the corpus with an existing text-to-AMR system for English. Because the corpus is aligned on the sentence level, this means that the same AMR of an English sentence, is also compatible with the same sentence in the other languages. The resulting, domain-specific, synthetic dataset is not publicly available.

The annotation efforts above are noteworthy and have had a positive impact on the field. However, on the one hand deep learning experiments often require a significantly larger dataset than the manual annotations in Spanish have provided so far, and on the other hand one may prefer general-domain AMR annotations over domain-specific

ones for broad applicability. An AMR dataset for Dutch simply does not exist yet.

3.2. AMR parsing

In research on automated text-to-AMR parsing, most work has focused on English – which in part can be attributed to the availability of large corpora, suitable for machine learning, such as the AMR 2.0 and 3.0 corpora described above ([Knight et al., 2017, 2020](#)). Performance of automated systems has increased markedly in the last years thanks to innovations such as the Transformer architecture ([Vaswani et al., 2017](#)), transfer learning where a pretrained language model is finetuned on the task of AMR parsing, and the use of automatically created, synthetic data for training (also called “silver” data in contrast to manually created “gold” data).

[Bevilacqua et al. \(2021\)](#), for instance, presented SPRING, a text-to-AMR and AMR-to-text model in **English** that was finetuned on a pretrained BART model ([Lewis et al., 2020](#)), outperforming previous approaches. They also showed that, in their set up, incorporating silver data did not positively affect the system’s performance. Following up on that, [Bai et al. \(2022\)](#) went a step further by also exploring pretraining a unified model in all directions: text-to-AMR, AMR-to-text, text-to-text, and AMR-to-AMR for English. Similarly, [Cheng et al. \(2022\)](#) proposed to unify AMR-to-text and text-to-AMR tasks but instead of using silver data they employed Bayesian multi-task learning. Also within the Bayesian paradigm, researchers at IBM ([Lee et al., 2022](#)) suggested that relying on self-supervised training with silver data in itself is not sufficient to push parsers’ performance higher anymore. In addition, they suggest the use of ensembling multiple system outputs together in combination with distillation for improved performance and efficiency. Noteworthy here is that they also apply their findings on Chinese, German, Italian and Spanish models where they set a new state-of-the-art on the “Four Translation” dataset. Their work relies heavily on earlier findings of [Zhou et al. \(2021\)](#), who explicitly integrated structural information of the AMR graph into pretrained language models. In a similar vein, most recently, [Vasylenko et al. \(2023\)](#) also modify the aforementioned Transformer architecture with adapters that are tailored to contain structural graph information, achieving state-of-the-art results as a non-ensemble system through distillation without the use of additional data.

To the best of our knowledge language-specific models for general-purpose **Dutch**-to-AMR parsing do not exist. Prior work has been done on semantic parsing for Dutch, but as noted in [Wang and Bos \(2022\)](#), no annotated AMR corpora exist for Dutch, so research for Dutch is focused on other

⁴AMR 2.0: <https://catalog.ldc.upenn.edu/LDC2017T10>; AMR 3.0: <https://catalog.ldc.upenn.edu/LDC2020T02>

semantic paradigms, such as Discourse Representation Graphs in multilingual settings (Wang et al., 2023). As mentioned before, some datasets have been created for **Spanish** AMR, but they are relatively small in size for extensive deep learning experimentation or they are not publicly available (Fan and Gardent, 2020), which leads to little research in text-to-AMR for Spanish specifically except for the work that was already referred to above due to their data creation efforts and broader research on multilingual systems.

In the aforementioned work of Lee et al. (2022), **multilinguality** is achieved through the use of machine translation as a data augmentation technique. This is common practice in other research as well in an attempt to automatically create sizable AMR corpora. Mitreska et al. (2022), for instance, establish text-to-AMR and AMR-to-text pipelines for Macedonian, German, Italian, Spanish and Bulgarian. The AMR parsing and generation itself is tailored to English, but they then use machine translation to translate the input or output to the relevant language. Using machine translation to translate English sentences while keeping the same AMR to create synthetic AMR data for other languages has been introduced and proved effective since Damonte and Cohen (2018), who showed a significant boost of performance in their multilingual AMR parsing when using machine-translated data.

While prominent in its descriptive nature for linguistic purposes, AMR’s increase of utility should also be mentioned. In the past year, AMR has been applied to NLP tasks ranging from machine translation (Song et al., 2019; Li and Flanagan, 2022) to the realm of multimodal research on the meaning and representation of gestures (Brutti et al., 2022) and images (Abdelsalam et al., 2022). The recent interdisciplinary endeavours underscore the broad exploration of AMR’s applicability. However, the impediments of data scarcity across various languages and the absence of automated systems in non-English linguistic domains pose substantial barriers to the advancement of research in this field.

4. Methodology

4.1. Model

In this work, all our models are finetuned from the same base model mBART (Liu et al., 2020), specifically its checkpoint `mbart-large-cc25`, to ensure a fair comparison. Note that despite the base model being multilingual for all our models, in our methodology we often refer to our “monolingual” and “multilingual” models to indicate we finetuned them. This Transformer-based (Vaswani et al.,

2017) encoder-decoder model was pretrained on the denoising objective of sequences (recovering an input text that had been scrambled, noised, deleted or otherwise modified) for 25 languages, including English, Spanish and Dutch. The data was resampled so that each language is equally represented in the training data of mBART. If we were to use different base models for each model, e.g. language-specific base models vs. multilingual base models, that would not be a fair comparison and it would not be clear whether the performance difference is caused by the amount of data or the quality of the base model. Therefore, for all of our models, we start from the same base model. Although we created Irish-Gaelic translations for other parts of our research, we did not include it in our model training. The reason is because mBART was not pretrained on Irish-Gaelic so the quality would not be fair compared to the other languages. We began working on this topic in mid-2022, but Heinecke and Shimorina (2022) demonstrated that the mT5 base model (Xue et al., 2021) is a suitable language model that covers Irish-Gaelic. We were not able to redo our work given computational and time constraints but hope that publishing our Irish-Gaelic data alongside the Spanish and Dutch variants enables other researchers to use the insights of Heinecke and Shimorina (2022) together with our data to create Irish-Gaelic AMR parsers.

Due to computational restrictions and because the base models are the same for all models (mBART), we did initial hyperparameter tuning for one model (`en+es+nl-part`) and its hyperparameters were then used to train other models as well. All models were trained for 25 epochs with early stopping.⁵ In the remainder of this paper we will make use of our translated AMR 3.0 dataset for training and evaluation.

Because mBART is a sequence-to-sequence model, our input (text) and output (AMR) data has to be formatted as a sequence of tokens. The graphs in the datasets are therefore linearised and delinearised back into a PENMAN representation with a reimplement of SPRING’s (de)linearisation methods (Bevilacqua et al., 2021). They suggest to linearise a graph in a depth-first manner by slightly modifying the PENMAN representation. An example of this process is given in Figure 1c. As much content as possible is retained, such as opening and closing brackets, relations, and concepts. However, instead of variable names they add special tokens to the vocabulary, called pointer tokens. Instance relationships are made implicit by removing the forward slash (/). Concepts and relationships are also added to the vocabulary explicitly instead of rely-

⁵Exact hyperparameters will be given in an appendix in the camera-ready version.

ing on the model’s subword tokenizer to ensure that the model learns about those tokens explicitly. When delinearising a model’s prediction back into a graph, SPRING uses an iterative graph restoration method to fix potential issues if the predicted tokens could not be readily reconstructed into a graph, which they show works robustly.

4.2. Data

A valuable contribution of our work is the parallel, multilingual dataset that we provide for Spanish, Dutch and Irish Gaelic (Irish not used in this paper). We base our data collection methodology on the premise that “AMR annotations can be successfully shared across languages” (Damonte and Cohen, 2018, p. 1147). Unlike Damonte and Cohen (2020), who translated a relatively small portion of the AMR 2.0 corpus, we employ the more extensive AMR 3.0 corpus (Knight et al., 2020) and automatically translate *all* partitions (train, development, test) to make it usable for deep learning experiments. This corpus comprises 59,255 parallel AMR structures and English sentences, partitioned into canonical training (55,635), development (1,722), and test (1,898) sets. Unlike the domain-specific Europarl corpus used by Fan and Gardent (2020), AMR 3.0 spans a wider array of domains, including discussion forums, Wikipedia, news broadcasts, and literary works. For translation, we opted for Google Translate API v3, which was consulted on September 11th, 2023.⁶ All 59,255 sentences were translated into Dutch, Spanish, and Irish Gaelic and manually corrected with regard to formal issues such as unexpected white-spaces or wrongly encoded characters. Similar to previous works mentioned above the AMR side remains unchanged for all languages - only the English source text was automatically translated. This process yields a large, parallel multilingual corpus with aligned AMR annotations that is sufficiently large and diverse in domain for multilingual machine learning experimentation. We make this dataset available with the same license as the original AMR 3.0 corpus on the LDC website under the name “AMR 3.0 - Dutch, Irish, and Spanish Machine Translations”.⁷

As described before, the goal of this study is to gauge the performance of multilingual systems compared to their monolingual counter-parts, paying particular attention to the amount of data per-

⁶A sample of Dutch translations was manually verified for quality. For the other languages, we specifically selected Google Translate for its high-quality translations, corroborated by the report, “The State of Machine Translation, 2023” (<https://intn.to/machine-translation-report-2023/>).

⁷The data submission is accepted by the LDC and its release is planned for the second part of 2024.

language that the model is trained on. To do so we train monolingual models for English, Spanish and Dutch as the baselines, where each model is trained on their respective full dataset of 55,635 training instances. We also train multilingual models, with English, Spanish and Dutch, and with only Spanish and Dutch. We are mostly interested in the multilingual models that were trained on a subset of the data so that the multilingual model has seen the same number of training samples in total as the monolingual models but distributed across languages. Furthermore, for reference, we also train multilingual models that are trained on the full dataset for each language to see how well multilingual models fare. This data distribution and corresponding models has been illustrated in Table 1. We thus control our training strictly on data size: the baseline models are trained on their full, monolingual dataset (**-only*), the partial multilingual models (**-part*) are only trained on a subset per language, and the full multilingual models (**-full*) are trained on the full dataset of all languages combined. The hypothesis is that the baseline, monolingual models will perform better than the full multilingual models, which in turn will perform better than the partial multilingual models. A small difference would justify the compute efficient (one multilingual model) and data efficient (multilingual model trained on partial datasets) utility of multilingual AMR parsing.

| model \ lang. | en | es | nl |
|---------------|--------|--------|--------|
| en-only | 55 635 | 0 | 0 |
| es-only | 0 | 55 635 | 0 |
| nl-only | 0 | 0 | 55 635 |
| en+es+nl-part | 18 545 | 18 545 | 18 545 |
| es+nl-part | 0 | 27 818 | 27 817 |
| en+es+nl-full | 55 635 | 55 635 | 55 635 |
| es+nl-full | 0 | 55 635 | 55 635 |

Table 1: Contents of the training set for each model. The row sections represent monolingual models (**-only*), multilingual models that have been trained only on part of the data per language (**-part*), and multilingual models that were trained on the full dataset for each language (**-full*).

4.3. Evaluation

For text-to-AMR parsing, it is common to use Smatch scores (Cai and Knight, 2013), which calculate the precision, recall and Smatch F1 scores on matching the triples of the predicted graph with the reference graph. We report Smatch F1 scores as calculated by `smatchpp` (Opitz, 2023), particularly its ILP solver rather than the hill climber ap-

proach for the best result. We report pairwise significance levels on the differences between systems based on this F1 score, by sorting systems best to worst and bootstrapping ($n = 1000$).⁸ For brevity we show compact tables in the paper that only contain smatch f1 scores and, for the coarse-grained results, their significance compared to the lower performing systems. In addition, we discuss more fine-grained evaluation scores that are common in AMR research, following Damonte et al. (2017). The following categories are reported:

- Unlabeled: Smatch score without considering the edge labels (the relation between two items)
- No WSD: Smatch score without word sense disambiguation (g_0 instead of g_0-01)
- Concepts: score of correctly predicting concepts
- Named entities: score of correctly predicting named entities (`:name`)
- Negations: score of correctly predicting negations and polarity (`:polarity`)
- Wiki: score of correctly predicting linked Wikipedia entries (`:wiki`)
- Reentrancy: some nodes can be reentering, for instance due to coreference (so they have more than one parent; like `b` in Fig. 1a).
- SRL: Smatch score for semantic role labelling, i.e., only considering `ARGn` relations to identify predicate-argument constructions

5. Results

In Tables 2, 4, and 6, we provide for all systems their smatch F1, precision and recall scores. All tables are sorted from worst to best according to the smatch F1 score. For each system the significance compared to only the previous system above it is given for conciseness reasons; other important significant differences as well as overlapping confidence intervals are described in the text. Bold fonts indicate best systems for a given metric. Note that we trained a multilingual model on Spanish and Dutch only to see whether leaving out English as a high-resource language would impact the results for the other languages. Therefore, the English results contain fewer systems than the other two languages.

Detailed scores on specific categories are given in Tables 3, 5, and 7, for English, Spanish and Dutch respectively. Here we report only the F1 scores (multiplied by 100). Digits after the decimal points are not reported for the fine-grained analysis due to the limited decimal precision in the fine-grained evaluation framework.

⁸<https://github.com/mdtux89/amr-evaluation/>

| | smatch f1 | smatch p | smatch r |
|---------------|----------------|--------------|--------------|
| en+es+nl-full | 79.07 | 79.92 | 78.24 |
| en+es+nl-part | 80.14** | 81.52 | 78.81 |
| en-only | 81.30** | 82.34 | 80.29 |

Significant differences with the previous row are marked as ** $p < 0.01$

Table 2: Smatch F1, precision and recall scores on the English test set

| | en+es+nl-full | en+es+nl-part | en-only |
|----------------|---------------|---------------|-----------|
| unlabeled_f | 82 | 83 | 84 |
| no_wsd_f | 79 | 81 | 82 |
| concepts_f | 85 | 87 | 88 |
| ner_f | 84 | 85 | 85 |
| negations_f | 63 | 66 | 69 |
| wiki_f | 74 | 74 | 75 |
| reentrancies_f | 68 | 69 | 71 |
| srl_f | 78 | 79 | 80 |

Table 3: Fine-grained evaluation results for the English test set (F1 score only)

| | smatch f1 | smatch p | smatch r |
|---------------|--------------|--------------|--------------|
| en+es+nl-part | 73.04 | 74.59 | 71.56 |
| es+nl-part | 73.36* | 74.76 | 72.79 |
| es+nl-full | 73.99 | 74.97 | 73.02 |
| en+es+nl-full | 74.10 | 74.99 | 73.24 |
| es-only | 74.56 | 75.85 | 73.30 |

Significant differences with the previous row are marked as * $p < 0.05$

Table 4: Smatch F1, precision and recall scores on the Spanish test set

| | en+es+nl part | es+nl part | es+nl full | en+es+nl full | es only |
|----------------|---------------|------------|------------|---------------|-----------|
| unlabeled_f | 77 | 78 | 78 | 78 | 78 |
| no_wsd_f | 73 | 74 | 74 | 74 | 75 |
| concepts_f | 76 | 77 | 77 | 77 | 78 |
| ner_f | 83 | 83 | 83 | 83 | 84 |
| negations_f | 52 | 58 | 55 | 55 | 59 |
| wiki_f | 71 | 72 | 73 | 73 | 73 |
| reentrancies_f | 62 | 62 | 62 | 62 | 63 |
| srl_f | 70 | 71 | 71 | 71 | 72 |

Table 5: Fine-grained evaluation results for the Spanish test set (F1 score only)

| | smatch f1 | smatch p | smatch r |
|---------------|--------------|--------------|--------------|
| es+nl-part | 73.09 | 74.15 | 72.07 |
| en+es+nl-part | 73.37 | 74.72 | 72.07 |
| en+es+nl-full | 73.45 | 74.24 | 72.66 |
| es+nl-full | 74.07 | 74.92 | 73.24 |
| nl-only | 74.36 | 75.60 | 73.15 |

No significant differences between successive rows

Table 6: Smatch F1, precision and recall scores on the Dutch test set

| | es+nl part | en+es+nl part | en+es+nl full | es+nl full | nl only |
|----------------|---------------|------------------|------------------|---------------|------------|
| unlabeled_f | 77 | 77 | 77 | 78 | 78 |
| no_wsd_f | 73 | 73 | 73 | 74 | 74 |
| concepts_f | 76 | 76 | 76 | 77 | 78 |
| ner_f | 82 | 84 | 84 | 84 | 84 |
| negations_f | 51 | 54 | 52 | 54 | 57 |
| wiki_f | 72 | 72 | 73 | 73 | 73 |
| reentrancies_f | 60 | 61 | 61 | 62 | 62 |
| srl_f | 69 | 70 | 70 | 71 | 71 |

Table 7: Fine-grained evaluation results for the Dutch test set (F1 score only)

6. Discussion

The central hypothesis of this study posited that monolingual systems *en-only*, *es-only*, and *nl-only* would outperform multilingual systems in terms of F1 Smatch scores. The empirical data affirm this hypothesis, revealing a consistent pattern where monolingual models surpass their multilingual counterparts across all three languages investigated. However, a nuanced interpretation of the statistical significance tests offers some promising insights.

English For coarse-grained results on English (Table 2), the monolingual model was found to be significantly better than both multilingual models. In absolute terms, however, this difference is small: the difference between the multilingual model that was only trained on part of the dataset for each language, and the monolingual model is only 1.2 Smatch F1, and on top of that their confidence intervals overlap. Interestingly, training on the full datasets with all languages combined yields significantly worse performance. This seems to indicate that for English, training on more non-English data deteriorates performance. This is unexpected because the assumption is that training on more data as a whole should yield better results, but given the significant difference between *en+es+nl-full* and *en+es+nl-part* that is not the case for English, i.e., added languages to an English dataset make results significantly worse regardless of the size of the data.

Digging deeper in the English results in Table 3, we find that there is a relatively small increase in scores across categories for each model, with the exception of the “negations” category, where a larger differences can be noted between all models. Negation, or rather the “polarity” of an utterance, has been proven difficult for automatic AMR parsers in earlier work, so much so that it has been suggested to post-process the AMR graph with a heuristic algorithms to re-apply negation based on polarity words in the input (Zhang et al., 2019). Such methods can positively impact performance; however, in this study we are interested in the effect of different data distributions on training re-

sults without any other modifications. In terms of negation, we see that mixing in other languages has a strong, negative impact.

Spanish Looking at the main results for Spanish (Table 5), the story changes in some respect. The differences between the monolingual model on the one hand and the full multilingual model *en+es+nl-full* and partial *en+es+nl-part* on the other are significant, with a difference in score of only 0.5 and 1.6 respectively. It is clear that the difference between the Spanish monolingual model and the full multilingual model of 0.5 is small and their confidence intervals overlap greatly. This is in sharp contrast with English, where – even though there also was small overlap between confidence intervals – the difference in Smatch score was larger with 1.2. Unlike English as well we see that the multilingual model trained on partial datasets performs significantly worse than all other models, including the multilingual model trained on all data. So unlike for English, training on full datasets with a lot of data from different languages improves the result, which was expected because that means the model has “seen” more diverse Spanish data as a whole. Scrutinising the bilingual models that were trained on only Spanish and Dutch, we find that the performance between them does not differ significantly. In fact, neither of them differ significantly from the second best performing system, the multilingual model trained on the full datasets *en+es+nl-full*. This sentiment is compounded when looking at the monolingual, best model and the worst bilingual model that was trained only on part of the data. While these models differ significantly, the difference is 1.2 Smatch and their confidence intervals overlap. The bilingual model trained on full datasets does not differ significantly from the monolingual model. So for Spanish, multi/bilingual models trained on the full dataset are viable. Furthermore, while the differences between the partial models and the monolingual one are significant, their differences are relatively small (1.6 and 1.2), and the confidence interval of the Spanish-Dutch model overlap with the one of the monolingual model, which indicates that training on non-English languages together with a Germanic language (Dutch) in limited data availability still yields good results that may be sufficient under data and compute constraints.

In the fine-grained results (Table 5), we see the same tendencies as for English. For all categories there is a slight increase in scores across systems. In many cases scores are even identical across systems, such as for all but the worst system for the unlabeled category, which indicates that the models are all similarly good at predicting the structure of the AMR graph and that a dif-

ference in performance is therefore mostly linked to how well they can predict the relations between nodes. Noteworthy, again, is the large difference in how well negations can be predicted. The monolingual model greatly outperforms the multilingual models in this respect but also the bilingual model `es+nl-part` performs well compared to the others, indicating that training on balanced, partial datasets *without* English seems to work well.

Dutch In Dutch we observe similarities with Spanish (Table 6). Multi/bilingual models trained on only a portion of the data perform worse than the monolingual model but absolute differences are small as we hypothesised: the gap between the worst and best model is only 1.3 Smatch. Whereas for Spanish the monolingual model did not differ significantly from the full multi/bilingual models, the monolingual model does differ from `en+es+nl-full` significantly, but only with $p = 0.046$, an absolute difference of 0.9 Smatch F1, and overlapping confidence intervals. Going back to the main interest, the models trained on partial data sets, we find that while the `-part` models differ significantly from the monolingual model (as expected) this is only 1.3 and 1 Smatch F1 respectively and in both cases the confidence intervals overlap. This indicates that for Dutch, training on partial datasets, even combined with a Romance language, yields competitive results compared to a monolingual model.

Dutch fine-grained results are consistent with our earlier findings (Table 7). Performance across categories is similar across all systems with the exception of negations. There, the monolingual model is again greatly outperforming the other systems. However, whereas `es+nl-part` yielded good results in the negation category for Spanish, it performs poorly in this category for Dutch.

7. Conclusion

Our findings suggest that our hypothesis is partially confirmed. For non-English languages, multilingual and even bilingual models achieve good quality. The gap between the worst and best model is 2.2 Smatch F1 for English, but only 1.6 for Spanish and 1.3 for Dutch. If annotated data is scarce for a language, or computational resources are limited to train or deploy multiple language-specific models, it is viable to instead train a single multilingual model with a small trade-off in performance.

Interestingly and unexpectedly, for English, adding too much data of other languages deteriorates model performance. A potential explanation might be that AMR concepts correspond to an English lemmas and training on a mix of plenty of

non-English and English data might “confuse” the model.

For all models and languages we confirm the findings of other researchers that polarity prediction is a hard task. We note that this category of errors alone seems to greatly impact performance across all models: in most of the fine-grained categories the performance difference between models is small but for “negations” it is fairly large. Therefore, using techniques such as the post-processing polarity algorithm by Zhang et al. (2019) could close the gap between multilingual and monolingual models even further.

By publishing our detailed findings, our models as baseline references, our multilingual dataset, and our training code, we hope to catalyse additional research in multilingual AMR parsing.

8. Limitations

Our work provides tangible language resources in the form of a multilingual AMR dataset and text-to-AMR models, and also offers insights into advantages and disadvantages of less-resource multilingual models. However, we also acknowledge limitations of our work.

To create our dataset, we make use of Google Translate, one of the best commercial MT systems available. However, we did not post-edit the translations or verified their translation in detail. Secondly, in our study we contrasted full monolingual models with partial and full multilingual models. In this study we did not include additional configurations, such as monolingual models with a subset of the dataset, or other data quantity variations. These were not feasible for us in terms of compute and time but could provide useful insights. Finally, we have based our methodology of training models mostly on the work of Bevilacqua et al. (2021). We have not made use of more recent work, nor used techniques such as multi-task learning, Bayesian learning or distillation. The impact of all those techniques on multilingual AMR parsing with machine-translated data could be promising.

By providing our models as a baseline alongside a multilingual dataset and training code, we aim to engage additional research that addresses these limitations that were out of scope for the current paper but that are noteworthy to investigate further.

Acknowledgements

This research was carried out as part of the Horizon 2020 “SignON” project, Grant Agreement No. 101017255. Model training and data processing was completed with the Flemish Supercomputer Center on grant “Extracting Meaning from Multilingual Text with Deep Neural Networks”.

9. Bibliographical References

- Mohamed Ashraf Abdelsalam, Zhan Shi, Federico Fancellu, Kalliopi Basioti, Dhaivat J. Bhatt, vladimir pavlovic, and Afsaneh Fazly. 2022. [Visual Semantic Parsing: From Images to Abstract Meaning Representation](#).
- Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. [Graph pre-training for AMR parsing and generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. [Abstract Meaning Representation for Sembanking](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Michele Bevilacqua, Rexhina Biloshmi, and Roberto Navigli. 2021. [One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12564–12573.
- Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. [Abstract Meaning Representation for Gesture](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583, Marseille, France. European Language Resources Association.
- Shu Cai and Kevin Knight. 2013. [Smatch: an evaluation metric for semantic feature structures](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.
- Ziming Cheng, Zuchao Li, and Hai Zhao. 2022. [BiBL: AMR parsing and generation with bidirectional Bayesian learning](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5461–5475, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Marco Damonte and Shay B. Cohen. 2018. [Cross-lingual Abstract Meaning Representation parsing](#). 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 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.
- Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. [An incremental parser for Abstract Meaning Representation](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 536–546, Valencia, Spain. Association for Computational Linguistics.
- Angela Fan and Claire Gardent. 2020. [Multilingual AMR-to-Text Generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2889–2901, Online. Association for Computational Linguistics.
- Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, reprinted edition. Number 42 in *Studies in Linguistics and Philosophy*. Springer-Science+Business Media, B.V, Dordrecht.
- Paul Kingsbury and Martha Palmer. 2002. [From TreeBank to PropBank](#). In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).
- Young-Suk Lee, Ramón Astudillo, Hoang Thanh Lam, Tahira Naseem, Radu Florian, and Salim Roukos. 2022. [Maximum Bayes Smatch ensemble distillation for AMR parsing](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5379–5392, Seattle, United States. Association for Computational Linguistics.
- Young-Suk Lee, Ramón Fernandez Astudillo, Radu Florian, Tahira Naseem, and Salim Roukos. 2023. [AMR Parsing with Instruction Fine-tuned Pre-trained Language Models](#). <https://arxiv.org/abs/2304.12272v1>.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](#). In *Proceedings of the 58th Annual Meeting of*

- the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Changmao Li and Jeffrey Flanigan. 2022. [Improving Neural Machine Translation with the Abstract Meaning Representation by Combining Graph and Sequence Transformers](#). In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 12–21, Seattle, Washington. Association for Computational Linguistics.
- 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.
- Christian M. I. M. Matthiessen and John A. Bateman. 1991. *Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese*. Communication in Artificial Intelligence. Pinter, London.
- Maja Mitreska, Tashko Pavlov, Kostadin Mishev, and Monika Simjanoska. 2022. [xAMR: Cross-lingual AMR End-to-End Pipeline](#). In *Proceedings of the 3rd International Conference on Deep Learning Theory and Applications*, pages 132–139, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.
- Juri Opitz. 2023. [SMATCH++: Standardized and extended evaluation of semantic graphs](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1595–1607, Dubrovnik, Croatia. Association for Computational Linguistics.
- Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. [Semantic Neural Machine Translation Using AMR](#). *Transactions of the Association for Computational Linguistics*, 7:19–31.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS 2017*, pages 1–15, Long Beach, CA, USA.
- Pavlo Vasylenko, Pere Lluís Huguet Cabot, Abelardo Carlos Martínez Lorenzo, and Roberto Navigli. 2023. [Incorporating graph information in transformer-based AMR parsing](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1995–2011, Toronto, Canada. Association for Computational Linguistics.
- Chunliu Wang and Johan Bos. 2022. Comparing Neural Meaning-to-Text Approaches for Dutch. *Computational Linguistics in the Netherlands Journal*, 12:269–286.
- Chunliu Wang, Huiyuan Lai, Malvina Nissim, and Johan Bos. 2023. [Pre-Trained Language-Meaning Models for Multilingual Parsing and Generation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5586–5600, Toronto, Canada. Association for Computational Linguistics.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. [AMR Parsing as Sequence-to-Graph Transduction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 80–94, Florence, Italy. Association for Computational Linguistics.
- Jiawei Zhou, Tahira Naseem, Ramón Fernandez Astudillo, Young-Suk Lee, Radu Florian, and Salim Roukos. 2021. [Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based AMR parsing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

10. Language Resource References

- Marco Damonte and Shay Cohen. 2020. [Abstract Meaning Representation 2.0 - Four Translations](#).
- Johannes Heinecke and Anastasia Shimorina. 2022. [Multilingual Abstract Meaning Representation for Celtic languages](#). In *Proceedings of*

the 4th Celtic Language Technology Workshop within LREC2022, pages 1–6, Marseille, France. European Language Resources Association.

Kevin Knight, Bianca Badarau, Laura Banarescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O’Gorman, and Nathan Schneider. 2017. [Abstract Meaning Representation \(AMR\) Annotation Release 2.0](#).

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Tim O’Gorman, Martha Palmer, Nathan Schneider, and Madalina Bardocz. 2020. [Abstract Meaning Representation \(AMR\) Annotation Release 3.0](#).

Noelia Migueles-Abraira. 2017. [A study towards Spanish abstract meaning representation](#). Master’s thesis, University of the Basque Country, June.

Shira Wein, Lucia Donatelli, Ethan Ricker, Calvin Engstrom, Alex Nelson, Leonie Harter, and Nathan Schneider. 2022. [Spanish Abstract Meaning Representation: Annotation of a general corpus](#). In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

MoCCA: A Model of Comparative Concepts for Aligning Constructicons

Arthur Lorenzi¹, Peter Ljunglöf², Benjamin Lyngfelt², Tiago Timponi Torrent^{1,3},
William Croft⁴, Alexander Ziem⁵, Nina Böbel⁵, Linnéa Bäckström⁶,
Peter Uhrig⁷, Ely Matos¹

¹Federal University of Juiz de Fora, ²University of Gothenburg, ³Brazilian National Council for Scientific and Technological Development – CNPq, ⁴University of New Mexico, ⁵HHU Düsseldorf, ⁶Halmstad University, ⁷University of Erlangen-Nuremberg

arthur.lorenzi@estudante.ufjf.br

Abstract

This paper presents MoCCA, a Model of Comparative Concepts for Aligning Constructicons under development by a consortium of research groups building Constructicons of different languages including Brazilian Portuguese, English, German and Swedish. The Constructicons will be aligned by using comparative concepts (CCs) providing language-neutral definitions of linguistic properties. The CCs are drawn from typological research on grammatical categories and constructions, and from FrameNet frames, organized in a conceptual network. Language-specific constructions are linked to the CCs in accordance with general principles. MoCCA is organized into files of two types: a largely static CC Database file and multiple Linking files containing relations between constructions in a Constructicon and the CCs. Tools are planned to facilitate visualization of the CC network and linking of constructions to the CCs. All files and guidelines will be versioned, and a mechanism is set up to report cases where a language-specific construction cannot be easily linked to existing CCs.

Keywords: construction, constructicon, comparative concepts

1. Introduction

Constructicons are digital collections of construction descriptions, in the sense and spirit of Construction Grammar. There are now such Constructicons available or under development for at least half a dozen languages (Lyngfelt et al., 2018), and more on the way. Some of them are developed for NLP purposes and others for language pedagogy – or both (Borin & Lyngfelt, *forthc.*; Ziem et al., *forthc.*). While these Constructicons are for the most part designed as monolingual resources, there is also ongoing work towards cross-linguistic application, or *multilingual constructicography*.

Previous efforts in Constructicon alignment have been designed as bilingual or trilingual comparisons (Laviola, 2015; Bäckström, Lyngfelt & Sköldberg, 2014; Lyngfelt et al., 2018). By using a constructional approach with tools from lexicography, the authors could discern close equivalents for most of the constructions investigated. However, the results showed a bias towards the source language, which was English, regarding both formal and functional properties. Furthermore, the method turned out to be far too time consuming to be feasible on a large scale and for more languages. Hence, the conclusion of these experiments is that, rather than construction by construction comparison, multilingual Constructicon alignment requires a language-neutral base of comparison.

In parallel, a series of workshops in Düsseldorf, Germany and Gothenburg, Sweden have been organized aimed at discussing possible methodologies for aligning constructions from the Brazilian Portuguese, English, German and Swedish Constructicons. The choice of Berkeley FrameNet frames as a possible comparative variable was obvious, not least because some Constructicons already link constructions and frames (Boas, Lyngfelt & Torrent, 2019). But it is clear that frames only link

the meaning/function of constructions and the form could therefore not represent a comparative variable. So the choice also fell on the use of Croft's comparative concepts (Croft, 2022), even if this entails a new implementation for almost all Constructicons. In addition to the selected methodology, these workshops also focused on practical implementation; in particular, rules and processes were developed for aligning different constructions using comparative concepts and frames.

In this paper, we present one of the outcomes of these workshops – the analytical and technical guidelines for aligning constructions and Constructicons via MoCCA (Model of Comparative concepts for Constructicon Alignment). These guidelines are jointly developed and agreed upon by Constructicon-building teams (CBTs) henceforth referred to as the CBT consortium.

The overall idea of this enterprise is to connect constructions across and within languages using comparative concepts (CCs) as a shared base of comparison (Lyngfelt et al., 2022). The CCs provide language-neutral definitions of linguistic properties, and language-particular constructions may be linked to any and all CCs conforming to properties shared by the construction in question. Thereby the construction will also be connected to other constructions linked to the same CC.

The guidelines described here are primarily directed towards Constructicon-building Teams, but may of course be employed by any linguist wishing to connect or compare a particular set of constructions to other constructions within or across languages via comparative concepts.

2. Comparative Concepts

Comparative concepts are the linguistic concepts used as the basis of cross-linguistic comparison in typology, although they were given this name only

recently (Haspelmath, 2010). Comparative concepts have been defined in terms of function (semantics, pragmatics) since Greenberg (1963) and Keenan & Comrie (1977). More recently, Haspelmath (2010) argues that some comparative concepts can be defined at least partly in terms of cross-linguistically valid properties of morphosyntactic form. Croft (2016, 2022) argues that comparative concepts are either completely functional, or are hybrids combining properties of function and form.

The CCs used in MoCCA are of five types: constructions, strategies (see 3.2.1), semantic content, information packaging and frames. The first four types, described in Croft (2022), are based on language typology. MoCCA uses an extension of the set of CCs presented by Croft (2022). Constructions and strategies are hybrid CCs, that is, pairings of form and function (see below). Semantic content and information packaging are purely functional CCs. The fifth type consists of the set of semantic frames defined in the Berkeley FrameNet 1.7 data release, as described in Ruppenhofer et al. (2016). Names of particular CCs, of any type, are written in boldface; when needed the CC type will be indicated within parentheses after the name.

Note that the language-neutrally defined constructions employed as CCs, henceforth CC-constructions, are not to be confused with language-particular constructions, which will here be called L-constructions. An L-construction, for example the English Polarity Question Construction (*Are you coming with us?*), is generally defined as the pairing of a language-specific form with a particular function.

In cross-linguistic comparison, particular constructions such as the **polarity question construction (cxn)** are compared first based on function: the set of form-function pairings across languages that express a function such as polarity questions (Weissweiler et al. 2024). The polarity question function is defined as **propositional content (sem)** packaged as an **interrogative (inf)** to which the interlocutor is expected to confirm, amend or disconfirm (Bolinger, 1978).

Cross-linguistic comparison starts from function because languages vary considerably in their morphosyntactic form in expressing functions such as the polarity question function. Although forms are language-specific, certain general properties of form can be defined cross-linguistically, for example, word order, prosody, a question particle, a special verb form, and so on. Thus, the English Polarity Question L-construction is an instance of the CC-construction **polarity question construction (cxn)** and uses the strategies of **word order (str)**—specifically, Subject-Auxiliary inversion—and **prosody (str)**—specifically, final rise intonation.

L-constructions may be linked to one or more CCs, and related L-constructions may share some CCs but differ with respect to others. Thus, the CC links represent partial correspondences and should not be confused with equivalence. Also note that the CCs cannot cover all properties of all constructions in all languages. There will always be language-particular idiosyncrasies not covered by this alignment model.

In the remainder of this paper we present the analytical and technical guidelines for using MoCCA, as well as the methodology for reporting issues with the system.

3. Analytical Guidelines

The MoCCA analytical guidelines focus on the procedures and principles for associating comparative concepts, including frames, with constructions and, if applicable, construction elements (CEs). In this section, we present its first version (1.0).

3.1. The CC Network

The CCs, including frames, are presented as a database of related concepts. The Croftian CCs can be related using nine different relations. *Subtype*, *part*, *attribute*, *value* and *role/filler* are used for CCs of the same type. In addition, for construction CCs, there is a special part relation, *head*, for the head of the construction.

Strategies are related to construction in three different ways. All strategies are related to the construction whose form they describe with the *expression-of* relation. Two classes of strategies also make reference to another construction.

One class of strategies, recruitment strategies, recruit the form of a related construction. For example, the English Physical Sensation construction (*I have a headache/a cold/etc.*) recruits the form of the Presentational Possession construction (*I have a car*). The *recruited-from* relation links this strategy to its source construction.

In another class of strategies, the system of strategies, elements of the form of one construction are based on corresponding elements in another construction. For example, in the accusative alignment strategy, illustrated by transitive *She saw her* vs. intransitive *She was sleeping*, the form of the A (transitive subject) argument phrase is the same as the form of the S (intransitive subject) phrase, i.e. *she*, while the form of the P (transitive object) phrase is different, i.e. *her*. In this case, the intransitive construction serves as the model for the strategy for encoding the transitive construction's arguments (A=S and P≠S). The accusative alignment strategy therefore has a *modeled-on* relation to the intransitive construction.

Finally, the *function* relation is used to link semantic and information packaging CCs to CC-constructions. Frame CCs can be related via the *inheritance*, *subframe*, *perspective-on*, *precedes*, *using*, *causative-of*, *inchoative-of*, *metaphor* and *see-also* relations (for a more detailed explanation on FrameNet relations, see Ruppenhofer et al., 2016).

The CC network is a directed acyclic graph (DAG), with arcs between CCs of the same type always having the direction of *child* to *parent*, *part* to *whole*, *attribute* to CC, *value* to *attribute* and *role/filler* to CC. Relations between different CC types are always directed to CC-construction. When building these relations, clusters of the same CC type are analyzed as taxonomic trees (e.g. the modification construction tree). When all trees of all five different

CC types are combined via the *expression-of*, *recruited-from*, *modeled-on* and *function* relations, the network becomes a DAG.

It is also possible that even within a CC cluster, there may be multiple, alternative taxonomies. This stems from the fact that language phenomena can be analyzed in different ways and to represent it, CCs can have multiple parents. When possible, however, multiple parents are always avoided as a way to make the network clearer.

To avoid inconsistencies, a set of constraints were devised and implemented to validate the state of the network. These validators, among other things, check whether the CC type constraints for each relation are respected and that no CC node is isolated. A more specific constraint, considered for CC-sem and CC-inf, would be: if X is a *value* of Y and Y is an *attribute* of Z, then X must be a *subtype* of Z.

In its current state, the network consists of 2286 CCs (of which 1222 are frames) and 3547 relations between them. The 1672 relations between non-frame CCs are all new and were manually created for MoCCA.

Previous experience with the Global FrameNet Shared Annotation task (Torrent et al., 2018; Giouli et al., 2020) reveals that users will often find the need to expand the model by adding new CCs or revising the existing ones. This is not a choice teams will be allowed to make on the fly, since it would compromise the alignment. Nonetheless, mechanisms are proposed for dealing with cases where teams cannot find a perfectly matching CC for the construction or construction element under analysis (see Section 5).

3.2. Associating CCs with Constructions

Associating CCs with constructions requires consideration of both function and form of constructions. This is not a simple task. For example, English uses a special pronoun for reflexive meaning for all persons, e.g. *I cut myself*, while Brazilian Portuguese and Swedish use the ordinary transitive object pronouns for 1st and 2nd person: Swedish *Jag skar mig* 'I cut myself' (lit. *I cut me*).

In contrast, for reciprocal meaning English uses a special multi-word expression: *We saw each other*, but some verbs express reciprocal meaning without any special form: *We met*. Brazilian Portuguese uses the same pronouns that are used for reflexive meaning: *Nos vemos amanhã na cidade* 'We'll see each other downtown tomorrow'; or a special form *um ao outro*, with or without the reflexive pronoun. Swedish uses a special form different from its reflexive pronoun, not unlike English: *Tvillingarna avskyr varandra* 'The twins detest each other'; but in some cases use a special verb form in -s: *De träffas och talas vid* 'They meet (each other) and talk (to each other)'.
This many-to-many mapping between form and function across languages requires us to compare form and function partly independently.

The association between CCs and L-constructions is guided by the following four principles, which will be discussed in detail in the subsections below.

1. **Application:** Link L-constructions to both a CC-construction and one of the CC-construction's CC-strategies.

2. **Generality:** Link an L-construction to the CC-construction at the lowest relevant level of generality in the CC-construction taxonomy.

3. **Constructional inheritance:** If there are multiple L-constructions in a taxonomic hierarchy, link them to the CC-constructions in the CC-construction taxonomic hierarchy that most closely matches the L-construction hierarchy.

4. **Analytical targets:** Link L-construction CEs to corresponding CC-constructions where the latter exist, and to CC-strategies where the L-construction CE is introduced by a strategy.

3.2.1. Application

In linguistic typology, function is the primary basis for cross-linguistic comparison, because function can be compared directly across languages. A typology of reciprocal constructions analyzes the variation in form for the expression of reciprocal meaning. This is an onomasiological approach to the analysis of constructions. Hence, "reciprocal construction" in a cross-linguistic sense is any morphosyntactic form expressing a particular function in any language.

Morphosyntactic form is language-specific, but some morphosyntactic properties can be defined in cross-linguistic terms. Variation in form across languages can be classified into morphosyntactic *strategies*, such as special pronoun form or a special verb form (Keenan and Comrie, 1977; Croft, 2022).

Thus, an L-construction in a Construction for a specific language includes a specification of a particular function (in cross-linguistic terms, a CC-construction) and its language-specific morphosyntactic form (a CC-strategy). An L-construction should therefore be linked to both a CC-construction and a CC-strategy of that CC-construction in the network.

Construction grammarians often take a semasiological approach, examining different functions expressed by a language-specific morphosyntactic form, such as the special reflexive pronoun form being used for reciprocal meaning in Brazilian Portuguese. For cross-lingual construction alignment, these correspond to two distinct CC-constructions, where one of the constructions has recruited the form of the other (a recruitment strategy).

3.2.2. Generality

When choosing the CC to be associated with a construction or CE, the most specific one that is applicable should be used.

The CCs in the linking model are organized in a network. Thus, linking an L-construction to a CC also connects it to related CCs of different generality and, indirectly, to associated L-constructions. This feature

may somewhat compensate for differences in granularity between L-construction entries in different Constructicons, through a graph structure identifying the closest corresponding target construction. Linking at too high a level of generality, however, may overgenerate and create less accurate connections. Therefore, one should try to find the most specific CC possible for any given construction or CE.

Sometimes this means that the best solution is to link to two or more co-hyponym CCs rather than a single hyperonym. If a supertype CC captures more subtypes than the ones relevant for characterizing the construction under analysis, then, the relevant subtypes should be associated with the construction, instead of the supertype.

3.2.3. Constructional Inheritance

Language-particular Constructicons may also be organized in inheritance networks. When looking at the network of CCs, it is important to consider the CC-construction's degree of generality/specificity in relation to that of the L-construction under analysis in the Constructicon.

For Constructicons that model construction inheritance, CCs and frames should be associated only once in an inheritance chain, at the adequate level of generality. For example, if a Constructicon has a general construction for relative clauses and four other constructions for subtypes of relative clauses, the more general **relative clause construction (cxn)** CC should be associated to the more general L-construction, while the subtypes – such as **anaphoric head (cxn)** and **free relative clause constructions (cxn)** – should be associated to the daughter L-constructions.

3.2.4. Analytical Targets

In the proposed linking model, CCs can be associated with constructions, CEs or both. If the Constructicon in question models constructional constituency in a way that allows for direct association of CEs to CC-constructions in a *part* relation to the CC-construction linked to the whole L-construction, CEs can be manually associated; they cannot be automatically derived from the *part* relation. If not, all applicable CCs should be associated at the level of the construction.

If the CC-construction already has CCs of its parts, they can be linked to L-construction CEs. For example, in linking the English Adjective Modification Construction, illustrated by *very large turkey*, the CCs **adjective modification construction (cxn)**, **adjective attributive phrase (cxn)**, and **referent expression (cxn)** are applicable to the whole referring phrase, its modifier (*very large*) and its head (*turkey*) respectively.

It is important to distinguish when the CE is associated with a CC-construction in a *part* relation, or a part of the CC-strategy used by the L-construction. For example, the English Finite Complement Clause L-construction, as in *Sally said that she ate the leftovers*, has four language-specific CEs: the matrix complement-taking predicate or CTP (*said*), the matrix Subject argument phrase (*Sally*), the complementizer (*that*), and the

complement (*she ate the leftovers*). The matrix CTP and its dependent arguments and complement are CEs of the CC-construction **complement clause construction (cxn)**. The complementizer, on the other hand, is introduced by the CC-strategy **complementizer (str)**.

4. Technical Guidelines

This section of the guidelines aims to provide Constructicon-building teams (CBTs) with information on the requirements for implementing the alignment between Constructicons. They cover issues concerning the database format, tools and versioning.

4.1. Database format

Considering the need to preserve the autonomy of different CBTs on how to organize and manipulate their data, MoCCA is split into a main database file for CCs and language specific files. These files follow the “keep it simple” principle, i.e. they are easy to read, both by humans and computer algorithms, and contain only information relevant for the Constructicon alignment.

4.1.1 The CC Database File

The first file, referred to as CC Database File, comprises the set of comparative concepts agreed upon and provided by the consortium and their relations. Since the CCs are the main features used to align constructions from different projects, this file should be treated as somewhat static. Changes are expected, but should not be drastic or as fast as other data, as they can potentially change the alignment of all Constructicons. The main content of the file is its version and the CCs themselves. Each CC entry must contain a unique, persistent CC ID and the CC's type, name, definition and relations. To represent relations, each CC entry must include attributes for each relation type described in Section 3.1. These attributes must contain a list of CC IDs with which the CC relates via that type. If that list is empty, the attribute may be omitted. A YAML-like schema for this file looks like this (a ? indicates an optional field):

- CC Database File version
- List of CCs:
 - CC ID
 - CC Type
 - CC Name
 - CC Definition
 - Relation r_i : list of CC IDs ?

4.1.2 The Linking Database Files

The second file type in MoCCA stores the linking between a Constructicon and the comparative concepts. A Linking Database File uniquely identifies a Constructicon among all others and for that reason must contain an alphanumeric identifier for that Constructicon, its name, version and the ISO 639-3 code of its language. This file must contain an explicit indication of which CC Database File version was used for the linking process. Its main content is a list of L-constructions. Each L-construction entry must have an ID, name and description and a list of associated CC IDs. The L-construction ID does not

need to be universally unique, i.e., it only needs to be unique for the Constructicon in question. The list of CC IDs is restricted to IDs from the CC Database File in its version specified by the linking file.

When applicable, the L-construction entry should also specify the ID of its parent L-construction and a list of its CEs in the Constructicon. The required data for CE entries mirrors that of L-construction entries: CE ID, name and description, parent CE ID and a list of CC IDs. Optional data for a Linking File includes name and descriptions in English and up to 3 example constructs. A representation of this complete schema in a YAML-like format looks like this:

- Constructicon ID
- Constructicon Name
- Constructicon Version
- CC Database File Version
- MoCCA Guidelines version
- Language ID (ISO 639-3)
- List of constructions:
 - L-construction ID
 - L-construction Name
 - L-construction Name (en) ?
 - L-construction Description
 - L-construction Description (en) ?
 - Parent L-construction ID ?
 - List of CC IDs (linking to the CC file)
 - List of Examples [0-3] ?
 - List of L-construction CEs:
 - CE ID
 - CE Name
 - CE Name (en) ?
 - CE Description
 - CE Description (en) ?
 - Parent CE ID ?
 - List of CC IDs (linking to the CC file)

4.2. Tools

Relating constructions to comparative concepts is the only way in which data from different projects can be connected. To make this process easier, faster and inconsistency-free, it is possible to develop a linking tool. This can be a web interface or API with which users (*i.e.* Constructicon developers) can easily link their constructions and CEs to CCs from MoCCA.

Another possible useful tool could use the linked databases from all existing Constructicons to show how different constructions are related between different languages. This can be done via different kinds of visualizations of the underlying CC graphs.

4.3. Versioning

To increase compatibility and preserve the ability of projects to work at their own pace, all of the files and guidelines previously discussed need to be versioned. Every database or tool built based on the CC or the Linking Databases needs to explicitly include the version of those files that was used as part of the metadata. In the case of the Linking Database Files, the Constructicon projects are expected to update their version according to the changes made and also provide the version of the

analytical guidelines that were followed. When changing the CC Database File version used by a Linking Database File, documentation will be provided to guide the automatic or manual update to a new version, depending on the changes made to the CCs by the CBT consortium.

5. Reporting Issues

This section presents guidelines for situations where it is not possible to adhere to the four principles in Section 3.2 or any other problem arises. In such cases, teams should report an issue. This reporting can be done using GitHub's Issues system at the appropriate repository under our organization¹.

A special type of issue are situations where the L-construction will not appear to fit into the function defined for any of the CC-constructions in the CC network, or the strategies defined for the relevant CC-construction. To fix that, a new CC must be proposed.

In the case of an apparent missing CC-construction, the report should include a proposed name and definition, the semantic content and information packaging CCs that define the proposed CC-construction, and taxonomic and/or partonomic relations between the proposed CC-construction and existing CC-constructions in the network.

In the case of an apparent missing CC-strategy, the report should include the name and definition and the ID of the CC-construction which the CC-strategy expresses, what type of strategy is involved, and in the case of a system strategy or a recruitment strategy, what system the strategy is part of, and what the source construction for the recruitment strategy is.

For more general problems or issues with other principles, the report must describe the situation and if possible, propose a change. This description should indicate whether the issue pertains to one of the four principles in Section 3.2 and where it relates to, *i.e.* guidelines or one of the database files. The solution, if present, must propose a change to that part of MoCCA.

In all cases, the CBT consortium will analyze the issue report and implement the solution that best handles the case while minimizing the impact to the full alignment.

6. Conclusion

In this contribution, we have introduced a model for aligning Constructicons based on comparative concepts (Croft 2016, 2022). More than 2,000 CCs, including all FrameNet 1.7 frames, have been collected in the MoCCA database and are made available to the research community together with a set of guidelines that define the process of aligning constructions from different Constructicon projects. In addition to these analytical and technical guidelines, a process for reporting issues and suggesting amendments has been outlined.

¹ <https://github.com/comparative-concepts>

While the model presented here is aimed at Constructicon-building teams (CBTs), we encourage linguists working with other languages and constructions to also consider linking their resources to MoCCA as a way to compare their work to existing Constructicons. By expanding the pool of languages linked to the comparative concepts, we can further improve the comparison model.

MoCCA can also form the basis of schemas and guidelines for other types of annotations, both within specific languages and across languages. It is also a useful resource for projects working with interlingua or other computational methods that could leverage the network of CCs. Finally, it could also serve in any type of contrastive work or as a resource for teaching.

Acknowledgements

MoCCA builds upon the foundation of previous work and from the discussions held at the Constructicon Alignment Workshop 2022, in Gothenburg (<https://www.globalframenet.org/caw2022>). We are grateful for the exchanges with Elodie Winckel, Kyoko Ohara, Thomas Herbst and Valentina Zhukova.

Authors acknowledge the support of the Graduate Program in Linguistics at the Federal University of Juiz de Fora. Research collaboration leading to MoCCA was funded by CAPES/STINT grant 99999.009910/2014-00 and CAPES/PROBRAL grant 88887.144043/2017-00. Lorenzi's research was funded by CAPES PROBRAL PhD exchange grant 88887.628831/2021-00 and CAPES PROEX PhD Grant 88887.816228/2023-00. Torrent is an awardee of the CNPq Research Productivity Grant number 315749/2021-0.

References

Bäckström, L., Lyngfelt, B., & Sköldberg, E. (2014). Towards interlingual constructicography: on correspondence between constructicon resources for English and Swedish. *Constructions and Frames*, 6(1), 9-33.

Boas, H. C., Lyngfelt, B., & Torrent, T.T. (2019). Framing constructicography. *Lexicographica*, 35(2019), 41-85.

Bolinger, D. (1978). Yes—no questions are not alternative questions. In *Questions* (pp. 87-105). Dordrecht: Springer Netherlands.

Borin, L., Lyngfelt, B. (Forthcoming). Framenets and constructicons. *The Cambridge Handbook of Construction Grammar*, ed. by Mirjam Friend and Kiki Nikofofidou. Cambridge University Press.

Comrie, B. (2014). Noun Phrase Accessibility and Universal Grammar 1. In *Universal Grammar (RLE Linguistics A: General Linguistics)* (pp. 3-45). Routledge.

Croft, W. (2016). Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology*, 20(2), 377-393.

Croft, W. (2022). *Morphosyntax: constructions of the world's languages*. Cambridge University Press.

Giouli, V., Piliitsidou, V., & Christopoulos, H. (2020). Greek within the Global FrameNet Initiative: Challenges and Conclusions so far. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet* (pp. 48–55). European Language Resources Association.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2, 73-113.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3), 663-687.

Laviola, A. (2015). *Frames e Construções em Contraste: uma análise comparativa português-inglês no tangente à implementação do constructicons*. Master's thesis, Federal University of Juiz de Fora.

Lyngfelt, B., Borin, L., Ohara, K., & Torrent, T. T. (Eds.). (2018). *Constructicography: Constructicon development across languages* (Vol. 22). John Benjamins Publishing Company.

Lyngfelt, B., Torrent, T. T., Matos, E. E. S., & Bäckström, L. (2022). Comparative concepts as a resource for multilingual constructicography. *Valency and constructions. Perspectives on combining words* (Meijerbergs arkiv för svensk ordforskning, 46), 100-29. Göteborg: Meijerbergs institut för svensk etymologisk forskning, Göteborgs universitet.

Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R., & Scheffczyk, J. (2016). *FrameNet II: Extended theory and practice*. International Computer Science Institute.

Torrent, T. T., Ellsworth, M., Baker, C. F., & Matos, E. E. (2018). The multilingual FrameNet shared annotation task: a preliminary report. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (pp. 62-68). European Language Resources Association.

Weissweiler, L., Böbel, N., Guiller, K., Herrera, S., Scivetti, W., Lorenzi, A., Melnik, N., Bhatia, A., Schütze, H., Levin, L., Zeldes, A., Nivre, J., Croft, W., & Schneider, N. UCxn: Typologically Informed Annotation of Constructions Atop Universal Dependencies. arXiv. <https://doi.org/10.48550/arXiv.2403.17748>

Ziem A., Willich A., & Michel, S. (Eds.). (Forthcoming). *Constructing Constructicons*. John Benjamins.

ISO 24617-8 Applied: Insights from Multilingual Discourse Relations Annotation in English, Polish, and Portuguese

Aleksandra Tomaszewska¹, Purificação Silvano², António Leal², Evelin Amorim³

¹Institute of Computer Science, Polish Academy of Sciences

²University of Porto/ Centre for Linguistics of the University of Porto

³Institute for Systems and Computer Engineering, Technology and Science

aleksandra.tomaszewska@ipipan.waw.pl;
msilvano@letras.up.pt; jleal@letras.up.pt;
evelin.f.amorim@inesctec.pt

Abstract

The main objective of this study is to contribute to multilingual discourse research by employing ISO-24617 Part 8 (Semantic Relations in Discourse, Core Annotation Schema – DR-core) for annotating discourse relations. Centering around a parallel discourse relations corpus that includes English, Polish, and European Portuguese, we initiate one of the few ISO-based comparative analyses through a multilingual corpus that aligns discourse relations across these languages. In this paper, we discuss the project's contributions, including the annotated corpus, research findings, and statistics related to the use of discourse relations. The paper further discusses the challenges encountered in complying with the ISO standard, such as defining the scope of arguments and annotating specific relation types like Expansion. Our findings highlight the necessity for clearer definitions of certain discourse relations and more precise guidelines for argument spans, especially concerning the inclusion of connectives. Additionally, the study underscores the importance of ongoing collaborative efforts to broaden the inclusion of languages and more comprehensive datasets, with the objective of widening the reach of ISO-guided multilingual discourse research.

Keywords: ISO 24617-8, discourse relations, parallel corpora

1. Introduction

Discourse relations are connections linking the meaning conveyed by two or more situations in discourse, articulated either explicitly or implicitly. The ISO-24617-8 standard provides a structured approach for annotating these relations in texts across various languages and genres. It is designed for use in natural language corpora and serves as a reference model for automated techniques in basic discourse parsing, summarization, and other related applications (ISO, 2020).

Importantly, ISO-24617-8 has the potential to advance multilingual discourse studies by offering a universal analytical framework. Despite its utility, projects utilizing this standard, especially in multilingual contexts, are rare. Our research addresses this by applying ISO-24617-8 to a corpus comprising Polish, English, and European Portuguese. The aim is to examine the distribution of discourse relations in these languages, along with the challenges of applying the standard to such data.

This study was carried out within the Multilingual Discourse Annotation Initiative (MDAI), an emerging collaboration in multilingual discourse analysis between Polish and Portuguese scholars. The initiative adopts the ISO 24617-8 standard for its versatility across different languages and genres.

In this paper, we present the inaugural study conducted by our team. Our work encompasses the development of research materials, pilot annota-

tions on select samples, and a trilingual annotation approach, offering early insights into the nature of discourse relations. Moreover, we examine the challenges we faced, especially in complying with the ISO standard, which paves the way for further refinement of the standard and its possible extension to other languages. The subsequent sections present our accomplishments, annotation methodologies, and initial findings.

Our main contributions are as follows:

- Testing ISO-24617-8 in a trilingual corpus to enhance comparative analyses and support multilingual discourse annotation.
- Providing statistics on the use of discourse relations across the three languages.
- Identifying challenges in adhering to the ISO-24617-8 standard for discourse annotation.

The paper is organized into six sections. The first section introduces the subject and outlines the research rationale. The second section reviews related work in the field, setting the context for the research. In the third section, we present ISO 24617-8, discussing its relevance and application to our study. The fourth section describes the research methodology, including the data collection process and the methods employed. The fifth section discusses the results of the study. The paper concludes with the sixth section, where we provide final remarks and propose future work in this area.

2. Related Work

Discourse relations are meaning relations between discourse units essential to understanding discourse structure and explaining different linguistic problems. They integrate semantic and pragmatic theories such as Theory of Discourse Coherence (Hobbs, 1985), Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), Taxonomy of Coherence Relations (Sanders et al., 1992), and Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003). These theories differ along several aspects, namely discourse relations’:

- designations – coherence (Hobbs, 1985; Sanders et al., 1992) or rhetorical relations (Mann and Thompson, 1988; Asher and Lascarides, 2003);
- definitions – based on semantic (Hobbs, 1985; Asher and Lascarides, 2003), pragmatic criteria (Grosz and Sidner, 1986) or a combination of the two (Mann and Thompson, 1988);
- nature – descriptive and operational constructs (Asher and Lascarides, 2003) or cognitive entities (Sanders et al., 1992; Mann and Thompson, 1988);
- number – for most proposals, an open list;
- arguments – type: clauses, single sentences, nominalizations; events, states or entities; simple or composite; adjacency: adjacent (RST) or also non-adjacent (SDRT);
- relevance – nucleus/satellite (RST) or subordinating/coordinating relations (SDRT).

Deriving discourse relations has been the subject of extensive research. One of the most comprehensive and well-founded frameworks for this purpose is SDRT, which combines a detailed formalization of the elements involved in discourse interpretation with semantic and pragmatic constraints to infer discourse relations. According to SDRT, there are two types of information sources responsible for computing a given discourse relation: linguistic sources, such as the lexicon and compositional semantics, and non-linguistic sources, such as world knowledge and the cognitive state of the participants.

Discourse relations can either be implicit, not signaled linguistically, or explicit (Taboada and Das, 2013). Explicit discourse relations are identified through the presence of a linguistic marker, which could be a word (e.g., ‘because’ for EXPLANATION), a lexical expression (e.g., ‘with the purpose of’ for RESULT), tense/mood/aspect (e.g., sequence of Simple Pasts for NARRATION), or syntactic structure (e.g., relative clause for ELABORATION). These linguistic markers are known as ‘discourse relational devices’, ‘connectives’ (van Dijk, 1979), ‘discourse markers’ (Schiffrin, 1987), ‘cue-phrases’ (Asher

and Lascarides, 2003), or ‘relational signs’ (Das and Taboada, 2019). Discourse relational devices play a significant role in triggering discourse relations and have been extensively studied (Iruskieta et al., 2014; Das and Taboada, 2019). Different taxonomies and findings have been reported in the literature to annotate datasets.

Various annotated datasets, comprising different genres and languages (individual or parallel), have been created for discourse relation identification. Some examples of these datasets are the RST-DT English corpus (Carlson et al., 2003); Penn Discourse Treebank (PDTB) (Prasad et al., 2008); RST Spanish Treebank (RST-ST) (da Cunha et al., 2011); SDRT Annodis French corpus (Afantenos et al., 2012); TED multilingual discourse bank (TED-MDB) (English, German, Polish, Portuguese, Russian, Turkish) (Zeyrek et al., 2018). Most of these datasets identify discourse relations through the presence of a discourse marker, while only a few rely on other sources of information (Benamara and Taboada, 2015).

Annotation is mainly done manually, either by trained linguists or non-experts, with a small number of instances of assisted automatic/semi-automatic annotation. (e.g., Gecco (Lapshinova-Koltunski and Anna Kunz, 2014); French Discourse Treebank (FDTB1) (Abeillé et al., 2000)).

The abundance of different frameworks makes it difficult to compare annotated corpora within the same language or across languages. Proposals such as ISO (ISO, 2016) (Bunt, 2015; Prasad and Bunt, 2015; Bunt and Prasad, 2016) aim to create interoperable, language-agnostic annotation schemes to address this issue. These annotated datasets with discourse relations are vital to Natural Language Processing (NLP) applications such as automatic summarization and translation, information retrieval, sentiment analysis, and opinion mining (Webber et al., 2012).

3. ISO 24617-8

ISO 24617 - Language Resource Management – Semantic Annotation Framework (SemAF) is made up of various components that tackle distinct facets of semantic annotation, including referential, temporal, and semantic role labeling. SemAF offers comprehensive coverage of linguistic phenomena. Part 8 – Semantic Relations in Discourse, Core Annotation Schema (DR-core) – ISO 24617-8 (ISO, 2016) deals with the annotation of locally established discourse relations.

The primary aim of ISO 24617-8 is to provide an interoperable approach to local discourse relations annotation, facilitating mapping between existing frameworks (e.g., RST (Mann and Thompson, 1988), SDRT (Asher and Lascarides, 2003),

PDTB (Prasad et al., 2008)) while adhering to the principles of the Linguistic Annotation Framework (ISO, 2012). It is also designed to be applicable to any natural language.

The “low-level” discourse relations proposed by ISO 24617-8 link two arguments, which are defined based on semantic criteria rather than syntactic. Thus, an argument of a discourse relation is any situation (state, event, fact, proposition or dialogue act), regardless of whether it is expressed syntactically by, for example, a nominalization, a clause, a sentence, or a discourse segment. Regarding the extent and adjacency of argument spans, ISO 24617-8 is neutral.

Each argument of a discourse relation is assigned an interpretation or role. Some discourse relations present pairs of arguments with the same role, so they are symmetric. Other discourse relations, called “asymmetric”, assign different roles to each argument. Similarly to other frameworks, in ISO 24617-8, discourse relations are established between the two arguments regardless of the existence or absence of discourse markers.

Figures 9 and 9 in the Appendix present the set of asymmetric and symmetric discourse relations put forward by ISO 24617-8.

According to ISO 24617-8, discourse relations are not a closed set, and many questions still need further research. For example, more precise distinctions are necessary for certain discourse relations, and there is a need for coverage of language-specific features, especially typologically distinct ones.

To the best of our knowledge, ISO-24617-8 has not been widely used to annotate discourse relations. Silvano et al. (2022) and Silvano and Damova (2023) propose a taxonomy grounded on ISO 24617-8 with a plug-in to Part 2 about Dialogue acts (ISO, 2020). This taxonomy represents the semantic and pragmatic meaning of discourse markers across nine different languages in a parallel corpus. Silvano et al. (2023) present an annotated corpus (DRIPPS) with discourse relations. It contains 993 sentences with adverbial perfect participial clauses in four varieties of Portuguese (European, Brazilian, Mozambican, and Angolan) and British English. The sentences were extracted from online newspapers and annotated with discourse relations following the ISO 24617-8 framework. The authors also annotated several discourse relational devices, such as connectors, the tense of the verb of the main clause, and the aspectual types of both clauses, to determine which ones contribute to the discourse relations inference.

Another resource is the Polish Discourse Corpus (PDC). Originating from a previous project that annotated discourse connectives to study their roles in various relations (Heliasz and Ogrodniczuk, 2019),

it is the first corpus designed for Polish that conforms to ISO 24617-8, and is unique in its multi-genre content. Comprising 1,745 texts from the Polish Coreference Corpus (Ogrodniczuk et al., 2015), the PDC reflects the genre distribution of the National Corpus of Polish (Przepiórkowski et al., 2012). An evaluation of the ISO 24617-8 standard’s application to Polish data revealed some challenges, especially with the subjective interpretation and vague definitions of discourse relations, indicating a need for clearer guidelines (Żurowski et al., 2023). The project has achieved several milestones, including the identification of over 17,881 discourse relations. Additionally, an early version of an automatic parsing tool has been developed, adopting a sequence-tagging approach to provide an initial assessment of the complexity involved in parsing discourse relations in Polish texts (Ogrodniczuk et al., forthcoming).

4. Method and Materials

The subsequent sections detail the objectives and methodology of this study, including the development of research materials, test annotations on selected samples, and the trilingual annotation of a complete text.

4.1. Objectives

The research focuses on testing the ISO-24617-8 standard’s application in the context of multilingual discourse analysis, targeting a corpus of Polish, English, and European Portuguese. The standard is recognized for its potential as a comprehensive framework for analyzing discourse relations across various languages and genres. However, its practical deployment has been limited, especially in multilingual settings. The study seeks to address this gap by examining how the standard can be applied to a diverse linguistic dataset, aiming to uncover the distribution and utilization of discourse relations within these languages.

Another objective is pinpointing the challenges encountered in adhering to the ISO-24617-8 framework for discourse annotation. Identifying them is essential for suggesting potential adjustments or enhancements to the standard, thereby improving its applicability and effectiveness in future research.

4.2. Dataset

The Multilingual Discourse Annotation Initiative (MDAI) dataset currently features 60 TED talks in English, European Portuguese, and Polish. English serves as the pivot language. The decision to use TED talks¹ was based on their accessibility, which

¹<https://www.ted.com/talks>

makes publishing the annotated dataset possible. Additionally, some other TED texts were annotated with discourse relations using other frameworks (Zeyrek et al., 2018), allowing for future annotation comparisons.

Our TED talks selection process relied on three key criteria:

1. availability in all three target languages,
2. a narrative nature, and
3. a length of 600 to 800 words.

As part of our pilot study, we chose to transcribe “The History of the World According to Cats” TED talk².

4.3. Annotation methodology

The annotation process engaged a diverse group of participants, including two early-career researchers and two scholars with substantial academic backgrounds, all of whom shared an interest and expertise in discourse annotation. This collaborative effort laid the groundwork for examining the ISO 24617-8 standard’s applicability across different languages.

The initial phase of the study involved selecting texts for annotation, designing a pilot dataset, and establishing a shared digital workspace to facilitate joint annotation and discussions. A sample text from open-source data was chosen for test annotation, aimed at aligning the annotation methods with the ISO 24617-8 standard and identifying differences in annotation strategies, particularly between the Polish and European Portuguese annotators with experience from their individual teams.

During the initial (test) annotation phase, the annotators worked on the sample text to ensure methodological consistency with the standard and to uncover any discrepancies in their approaches. Following this, a meeting was convened to discuss and resolve differences, especially concerning argument length and content, definitions and categorizations of relations, and relation hierarchies. Both teams identified these aspects as challenging.

During the pilot review, minor discrepancies were addressed, particularly in argument scope and the distinction between certain relations. After thorough discussion, consensus was reached, with EXPANSION maintained as a crucial component of the ISO standard, despite the omission in the Polish Discourse Corpus (PDC) annotation (Żurowski et al., 2023), for instance. ELABORATION was defined to include instances where arguments pertained to the same event or situation, as stipulated

²The video is available at https://www.ted.com/talks/eva_maria_geigl_the_history_of_the_world_according_to_cats/transcript.

by ISO, but we also incorporated insights from SDRT’s interpretation of ELABORATION. For this reason, cases where the second argument portrayed a subevent of an event introduced by the first argument were also annotated as ELABORATION. Additionally, following Prévot et al. (2009), whenever the second argument provided more information about an entity represented in the first argument, the selected discourse relation would be ELABORATION as well.

The annotation proper was conducted on an entire transcription of a TED talk titled “The History of the World According to Cats” in English, European Portuguese and Polish. We prepared a working document with the rules of the MDAI Annotation Scheme, grounded in the ISO 24617-8 standard. The process included the identification of the text span, discourse connectives, argument scopes, determination of arguments’ order, identification of discourse relations, and arguments’ role. The English version was annotated by three non-native annotators, fluent in English, with expertise in discourse relations and experience with ISO 24617-8. The European Portuguese and Polish translations were each annotated by two native experts.

Following the annotation proper, we have conducted the subsequent parts of the study: (i) assessing inter-annotator agreement, (ii) discussing the findings, and (iii) challenges.

5. Findings and Discussion

In this section, we describe the results of our study. We begin by presenting the results regarding the different tasks, and then we elaborate on one of the tasks, discourse relations identification, discussing some of the challenges we faced.

5.1. Results

Text spans Table 1 reveals that while there is a general consensus among annotators within each language, there are discrepancies in the number of text spans identified across languages.

| | A1 | A2 | A3 | A4 |
|------------|----|----|----|----|
| English | 72 | 72 | 61 | - |
| Polish | - | - | 55 | 47 |
| Portuguese | 74 | 76 | - | - |

Table 1: The number of text spans identified in the three texts

Notably, the Polish subcorpus had fewer annotated text spans than the Portuguese dataset. The English and Portuguese datasets were annotated by the same individuals (A1 and A2), while A3 was among the annotators for the Polish subcorpus. It

is worth noting that all annotators had prior experience working with ISO 24617-8 in other projects. For this pilot study, they were provided with the guidelines presented by ISO 24617-8. The standard’s impartiality towards text span length may account for this variance. Moreover, these results point to some indefiniteness as to what an argument should be by some annotators, who seem to have a broader notion and do not conduct a finer-grained analysis of all possible arguments and the discourse relations between them.

For the inter-annotator agreement (IAA) of the span texts, we opted to follow a pairwise BLEU-1 approach due to the difficulties associated with measuring the traditional Cohen’s kappa in text span labeling (Deleger et al., 2012; Brandsen et al., 2020; Miranda, 2023). Some other scores to measure agreement are also possible. Carlson et al. (2003) mapped the hierarchical structures of the discourse into sets of units and then computed the Cohen’s kappa of the categorical sets, while Zeldes (2017) employed an automatic tagger to compare with the human annotations and then obtaining the accuracy between automatic and manual labels. However, grouping in a set of units makes the agreement score not intuitive to interpret since it is necessary to detail which groups exist and their proportions. The exact accuracy of spans can also not reflect the labeling work done, because to measure the argument agreements, we allowed some minor disagreement (up to 20% in the BLEU-1 score) in the text spans. Hence, the BLEU-1 score seemed a rational choice in the context of our research.

The BLEU score is a standard metric to evaluate the results of translation task (Papineni et al., 2002). The BLEU-1 is a variation of the BLEU score that considers the tokens in the reference and the target as one gram and computes the proportion of tokens from the target that appears in the reference. This score ranges from 0 to 1, where 0 is no match between the tokens of two texts, and 1 is the full match of the tokens of the reference and target texts. To compute the BLEU-1 of the annotated text spans, we consider one annotator as the reference, i.e., the gold standard, and the other annotator as the target. Then, we calculated the BLEU-1 score for each text span. If a text span does not present a match in the reference, then we set the BLEU-1 score of that annotation as 0. Next, we average the BLEU-1 scores of all text spans. After that, we switch the reference annotator and the target and compute the BLEU-1 score again. Finally, we average these two scores. Table 2 describes the agreement between annotators in each dataset.

Overall, the results indicate that identifying text spans, which may not necessarily be limited to minimal chunks, led to a reasonable level of agreement. However, it is worth noting that there was a different

| | $A_{1,2}$ | $A_{1,3}$ | $A_{2,3}$ | $A_{3,4}$ |
|------------|-----------|-----------|-----------|-----------|
| English | .63 | .65 | .48 | - |
| Polish | - | - | - | .63 |
| Portuguese | .67 | - | - | - |

Table 2: The IAA of the text spans as BLEU-1 score between annotators A1 and A2 ($A_{1,2}$), A1 and A3 ($A_{1,3}$), A2 and A3 ($A_{2,3}$) and A3 and A4 ($A_{3,4}$).

interpretation of the definition of the relevant text span for breaking down arguments by A3.

Example 1 illustrates some of the divergences observed in the annotations.

Example 1

He rode to Gibraltar with the rescued crew and served as a ship cat on three more vessels – one of which also sank.

The three annotators agree that the discourse relation ASYNCHRONY should link two situations. The initial situation, which has the argument role of Before, is "he rode to Gibraltar with the rescued crew", while the second situation, which has the role of After, is "(he) served as a ship cat on three more vessels". However, the annotators showed some disagreement regarding the extent of the second argument. A2 incorporated "one of which also sank" in the second argument ("and served as a ship cat on three more vessels – one of which also sank"), whereas A1 did not include this part ("and served as a ship cat on three more vessels"). Furthermore, A3 excluded the conjunction "and".

Arguments Another task we conducted during our pilot study was identifying the arguments for the selected text spans. For the IAA, we have only considered the cases where there was agreement of at least 0.8 in the BLEU-1 score on the text span. The agreement of arguments is computed in a similar way to the text spans agreement. Table 3 describes the BLEU-1 score for the arguments identified by the annotators.

The IAA for identifying arguments is higher compared to the IAA for identifying text spans. ISO 24617-8 has established clearer criteria for identifying arguments, which is not the case for identifying text spans. However, in certain cases, the absence of specific information can lead to inconsistent annotation of arguments. ISO 24617-8 defines an argument as an event, state, fact, proposition or dialogue act, but it does not address problematic cases, like the example 2.

Example 2

A população estava a aprender a dominar a natureza

The population was learning to dominate nature

| | $A_{1,2}$ | $A_{1,3}$ | $A_{2,3}$ | $A_{3,4}$ |
|------------|-----------|-----------|-----------|-----------|
| English | .83/.83 | .88/.88 | .76/.80 | - |
| Polish | - | - | - | .89/.82 |
| Portuguese | .84/.84 | - | - | - |

Table 3: The IAA agreement of the arguments (arg1/arg2) as BLEU-1 score between annotators A1 and A2 ($A_{1,2}$), A1 and A3 ($A_{1,3}$), A2 and A3 ($A_{2,3}$) and A3 and A4 ($A_{3,4}$).

One of the annotators identified two sentence fragments, "The population was learning" and "mastering nature", and connected them using the discourse relation SYNCHRONY. However, the other annotator believed that "to learn" and "to master" conveyed the same idea, so they only identified one sentence fragment.

Sometimes, there is disagreement because of how the connective is included in the sentence fragment. Various annotator teams have different practices; for instance, the team of Polish-language annotators treated connectives as a separate category and did not include them within sentence fragments. On the other hand, annotations from Portuguese annotators consistently show that connectives are always part of the second sentence fragment. The following examples demonstrate the discrepancy in how different annotators interpreted the same sentence fragment.

Example 3

(Arg 1) For the next several months this cat hunted rats and raised British morale (Arg 2) until a sudden torpedo strike shattered the hull and sank the ship. [Connective marked and included in Argument 2]

(Arg 1) For the next several months this cat hunted rats and raised British morale until (Arg 2) a sudden torpedo strike shattered the hull and sank the ship. [Connective marked and not included in Argument 2]

While this difference may slightly affect the argument span, it does not clearly lead to divergent interpretations of discourse relations.

ISO 24617-8 provides a flexible and neutral (core) framework, accommodating diverse interpretations of i.a., number of events in text spans. Each annotation project necessitates the development of tailored guidelines to adapt the ISO framework to its specific requirements, including addressing unique cases. Nonetheless, for enhanced interoperability, it would be better if these were addressed directly within the ISO standard, ensuring consistency and ease of application across different projects and languages.

Discourse relations Following the identification of the arguments, the annotators identified discourse relations. To compute the agreement of the discourse relations, we employed Cohen's kappa metric, which is a traditional way to analyze the

inter-rater reliability of categorical data (McHugh, 2012). Since the discourse relations comprise a set of classes, i.e. categorical data, we chose this methodology. Cohen's kappa values range from -1 to +1, where -1 represents total disagreement and +1 total agreement. Furthermore, when Cohen's kappa results in values around 0, then the amount of agreement expected is no more than what could occur by random chance. The IAA regarding the identification of discourse relations is presented in Table 4.

| | $A_{1,2}$ | $A_{1,3}$ | $A_{2,3}$ | $A_{3,4}$ |
|------------|-----------|-----------|-----------|-----------|
| English | .52 | .76 | .56 | - |
| Polish | - | - | - | .73 |
| Portuguese | .52 | - | - | - |

Table 4: The IAA of discourse relations as Cohen Kappa score between annotators A1 and A2 ($A_{1,2}$), A1 and A3 ($A_{1,3}$), A2 and A3 ($A_{2,3}$) and A3 and A4 ($A_{3,4}$).

Concerning the English text, the measurement of Cohen's kappa relative to A1/A2 and A2/A3 is moderate, while for A1/A3 is substantial. Within the same language, we observe different results. Cohen's kappa is moderate in Portuguese annotators, while it is substantial in the case of Polish annotators.

The identification of the argument role was the subsequent task of the annotators. Tables 5 and 6 present the IAA scores for identifying Argument 1 and Argument 2 roles.

| | $A_{1,2}$ | $A_{1,3}$ | $A_{2,3}$ | $A_{3,4}$ |
|------------|-----------|-----------|-----------|-----------|
| English | .95 | 1.0 | .94 | - |
| Polish | - | - | - | 1.0 |
| Portuguese | .92 | - | - | - |

Table 5: The IAA of Arg1 as Cohen Kappa score between annotators A1 and A2 ($A_{1,2}$), A1 and A3 ($A_{1,3}$), A2 and A3 ($A_{2,3}$) and A3 and A4 ($A_{3,4}$).

The identification of the arguments' role was for the most part consistent with the IAA scores reaching perfect agreement in the three languages and between all the annotators.

| | $A_{1,2}$ | $A_{1,3}$ | $A_{2,3}$ | $A_{3,4}$ |
|------------|-----------|-----------|-----------|-----------|
| English | .91 | .94 | 1.0 | - |
| Polish | - | - | - | 1.0 |
| Portuguese | .92 | - | - | - |

Table 6: The IAA of Arg2 as Cohen Kappa score between annotators A1 and A2 ($A_{1,2}$), A1 and A3 ($A_{1,3}$), A2 and A3 ($A_{2,3}$) and A3 and A4 ($A_{3,4}$).

5.2. Discourse Relations Identification: Challenges

We can draw some conclusions by zooming in on the results of the discourse relations’s annotation. The statistics in the table 7 rank the relations based on their prevalence across the three languages.

In terms of quantity, the Polish and Portuguese subcorpora showed little variation in the number of discourse relations identified (55 and 47 in Polish, and 76 and 74 in Portuguese), suggesting consistency within individual languages. In the English corpus, the counts were similar for two annotators (72 each) with similar annotation experience and lower for the third (61) from another team.

The analysis indicates that the agreement among annotators ranged from moderate to substantial, highlighting the variety in their interpretations. When reviewing the annotations across three languages, clear patterns emerged, especially in the frequency of certain discourse relations, suggesting a need for more specific discourse relations. Notably, the EXPANSION discourse relation exhibited significant variability, with counts of 22, 19, and 7 instances by different annotators within the English corpus. This variation points to different interpretations of this relation by the annotators, indicating an area for guideline improvement. In contrast, the CONCESSION and ELABORATION relations showed more consistency among annotators. For instance, in the English corpus, CONCESSION was marked 4 times by two annotators and 3 times by another, while ELABORATION was noted 5, 2, and 4 times, respectively. This suggests that the definitions for these relations might be clearer or more intuitive for the annotators. Relations such as CAUSE, ASYNCHRONY, and CONJUNCTION were annotated more frequently, possibly indicating clearer definitions or boundaries for these categories. This higher frequency could be due to the explicit nature of these relations, which often occur with connectives such as "and" in the case of CONJUNCTION or "because" in the case of CAUSE. Conversely, FUNCTIONAL DEPENDENCE, MANNER, and EXCEPTION were less commonly noted, and several discourse relations like EXEMPLIFICATION, CONDITION, NEGATIVE CONDITION, EXCLUSION, SUBSTITUTION, and FEEDBACK DEPENDENCE were not identified at all. This observation might relate to the dataset’s nature or

size but also may suggest a need to reassess the clarity and practicality of the definitions for these less frequently identified discourse relations. The initial analysis suggests that disagreements on annotated discourse relations often arise when an example can be interpreted according to the definitions of two distinct discourse relations. This underscores the nuanced nature of discourse relation annotation and highlights the need for more precise guidelines. Such is the case with example 4 from the Portuguese text:

Example 4

(Arg1) os gatos têm trabalhado lado a lado com os humanos há milhares de anos (Arg2) ajudando-nos, assim como nós os ajudamos

(Arg1) cats have worked side by side with humans for thousands of years (Arg2) helping us, just as we help them.

In this example, annotators agreed on the spans of both arguments and decided that Arg1 and Arg2 denoted the same situation. However, A1 identified ELABORATION, considering that Arg2 provides more detail about this situation than Arg1. A2 identified RESTATEMENT, clearly interpreting Arg2 from a different perspective.

One of the most significant differences between annotators concerns SYNCHRONY. In European Portuguese, A2 identified seven instances of SYNCHRONY, while A1 only identified three. The same annotators chose the same relations in the English subcorpus. A3 concurred with A1, identifying the feature in three instances. In the Polish subcorpus, each annotator recognized SYNCHRONY five times. The consistency observed in the Polish-language examples may stem from the explicit presence of connectives or cue phrases that indicate events occurring simultaneously, thereby easing the identification of this particular relation. The example presented in 5 illustrates this observation.

Example 5

(Arg 1) Oswojenie kota domowego miało miejsce 10 tysięcy lat temu na terenie starożytnego Bliskiego Wschodu wraz z (Arg 2) początkiem Neolitu.

(Arg 1) The domestication of the house cat took place 10 thousand years ago in the territory of the ancient Near East, together with (Arg 2) the beginning of the Neolithic period.

A different case may be observed in European Portuguese, illustrated by example 6.

Example 6

(Arg1) um gato preto e branco agarrado a uma tábua (Arg2) que flutuava

(Arg1) a black and white cat clinging to (Arg2) a floating board

Table 7: Comparative Annotation Frequencies Across Annotators for discourse relations.

| Discourse Relation | English | | | Portuguese | | Polish | |
|-----------------------|---------|----|----|------------|----|--------|----|
| | A1 | A2 | A3 | A1 | A2 | A3 | A4 |
| EXPANSION | 7 | 22 | 19 | 23 | 22 | 3 | 3 |
| ASYNCHRONY | 9 | 12 | 13 | 16 | 15 | 7 | 6 |
| CONJUNCTION | 11 | 8 | 7 | 7 | 7 | 13 | 11 |
| CAUSE | 9 | 8 | 12 | 11 | 8 | 9 | 8 |
| ELABORATION | 5 | 2 | 4 | 3 | 2 | 3 | 2 |
| CONCESSION | 4 | 3 | 4 | 4 | 4 | 4 | 4 |
| SYNCHRONY | 3 | 7 | 3 | 3 | 7 | 5 | 5 |
| CONTRAST | 2 | 4 | 1 | 2 | 4 | 2 | 1 |
| SIMILARITY | 3 | 1 | 2 | 2 | 1 | 0 | 0 |
| RESTATEMENT | 2 | 1 | 3 | 3 | 1 | 2 | 1 |
| MANNER | 2 | 0 | 0 | 0 | 0 | 1 | 1 |
| PURPOSE | 2 | 1 | 1 | 0 | 1 | 2 | 2 |
| EXCEPTION | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| FUNCTIONAL DEPENDENCE | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| DISJUNCTION | 0 | 0 | 0 | 0 | 0 | 3 | 2 |

In this case, A2 considered that Arg2 expanded on the setting relevant for interpreting Arg1 (EXPANSION), while A1 annotated SYNCHRONY. It is worth noting that temporal overlapping characterizes both SYNCHRONY and EXPANSION. A similar distinction in assigning temporal and non-temporal relations can be observed for Polish. One of the annotators uses CONJUNCTION for the discourse relation in example 7 whereas the other uses ASYNCHRONY for a similar instance with "oraz" (and), indicating a temporal sequence rather than a simple conjunction.

Example 7

(Arg 1) *Został ochrzczone Niezatapialnym Samem, popłynął na Gibraltarcz z ocalałymi członkami załogi oraz (Arg 2) pełnił służbę jako kot pokładowy na trzech innych okrętach*

(Arg 1) *He was named Unsinkable Sam, sailed to Gibraltar with the surviving crew members, and (Arg 2) served as a ship's cat on three other ships.*

In another example, one of the annotators interprets the use of *czy* (whether/or/ and) in the phrase *nie były chętne do kontaktu z innymi kotami czy ludźmi* (were not keen on contact with other cats or people) as indicating a DISJUNCTION, assigning the roles of "disjunction 1" and "disjunction 2". Conversely, another annotator views a similar usage of *czy* as an indicator for CONJUNCTION, thus labeling it with the roles "conjunction 1" and "conjunction 2", illustrating the variability in understanding the connective's function in discourse.

The following example is evidence of the complexity of the annotation and of how disagreement can occur. In Portuguese, as in other languages, the same verb can occur as main or auxiliary without morphological differences. The example 8 illustrates this feature.

Example 8

um contratorpedeiro inglês veio recolher os prisioneiros

an English destroyer came to collect the prisoners

In this case, A2 interpreted the sequence as denoting two distinct situations represented by two main verbs, Arg1 being "an English destroyer came" and Arg2 "to collect the prisoners", linked by the discourse relation PURPOSE. A1 annotated this text span as representing one situation, assigning to the verb "came" an auxiliary role, and for that reason, the discourse relation PURPOSE was not identified. Once again, although the guidelines established for each project can specify how to proceed in ambiguous cases, we argue that such instructions could be given by the ISO to allow for a more standardized approach.

6. Conclusions and Future Work

This study applied the ISO 24617-8 standard to a parallel corpus in English, Polish, and European Portuguese, aiming to explore the potential and challenges of using this framework for multilingual discourse analysis. The primary contribution is the annotated corpus, which offers insights into the use of discourse relations and connectives across the three languages.

During the initiative, we have encountered the challenge of operating without specific ISO-based guidelines for individual languages, prompting us to discuss and converge on collective interpretations. The DR-core, while foundational, presents moments of neutrality and ambiguity that required careful consideration. The annotation process was inherently time-consuming. Additionally, the

scarcity of existing multilingual discourse annotations emphasized the innovative aspect of our work, though it also meant we had no direct benchmarks for comparison.

Our analysis revealed varying interpretations and applications of the ISO standard, highlighting the need for more explicit guidelines, especially in defining the scope of arguments and categorizing specific types of relations. Transitioning from the challenges encountered, the outcomes of the project have so far been promising. The findings offer initial insights into the use and nature of discourse relations in the three languages, along with an analysis of the challenges encountered in adhering to the standard.

Future efforts will focus on expanding the corpus to include a broader range of languages and genres, which could help in understanding the universality and flexibility of the ISO standard in diverse linguistic contexts. Refining the annotation guidelines based on the experiences and challenges encountered in this study will be a priority, with an aim to improve the clarity and applicability of the ISO framework for discourse analysis as well as inter-annotator agreement.

Acknowledgements

This article is based upon work from COST Action NexusLinguarum³ — European network for Web-centered linguistic data science (CA 18209)⁴, supported by COST (European Cooperation in Science and Technology)⁵. The work was supported by the European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, CLARIN — Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00–00C002/19⁶, the Polish Ministry of Education and Science grant 2022/WK/09 and as part of the investment CLARIN ERIC — European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024–2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01. Portuguese national funds also funded this paper through FCT – Fundação para a Ciência e a Tecnologia, I.P., within the project UIDB/00022/2020.

³<https://nexuslinguarum.eu/>

⁴<https://www.cost.eu/actions/CA18209/>

⁵<https://www.cost.eu/>

⁶<https://clarin.biz/>

7. Bibliographical References

- Anne Abeillé, Lionel Clément, and Alexandra Kinyon. 2000. [Building a treebank for French](#). In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).
- Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA).
- Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, United States.
- Farah Benamara and Maite Taboada. 2015. [Mapping different rhetorical relation annotations: A proposal](#). In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.
- Alex Brandsen, Suzan Verberne, Milco Wansleben, and Karsten Lambers. 2020. [Creating a dataset for named entity recognition in the archaeology domain](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.
- Harry Bunt. 2015. [On the principles of semantic annotation](#). In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.
- Harry Bunt and Rashmi Prasad. 2016. ISO DR-core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. [Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory](#), pages 85–112. Springer Netherlands, Dordrecht.

- Iria da Cunha, Juan-Manuel Torres-Moreno, Gerardo Sierra, Luis-Adrián Cabrera-Diego, Brenda-Gabriela Castro-Rolón, and Juan-Miguel Roland Bartilotti. 2011. [The RST Spanish treebank on-line interface](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 698–703, Hissar, Bulgaria. Association for Computational Linguistics.
- Debopam Das and Maite Taboada. 2019. [Multiple signals of coherence relations](#). *Discours [En ligne]*, 24:1–38.
- Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, and Imre Solti. 2012. [Building gold standard corpora for medical natural language processing tasks](#). *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:144–153.
- Barbara J. Grosz and Candace L. Sidner. 1986. [Attention, intentions, and the structure of discourse](#). *Computational Linguistics*, 12(3):175–204.
- Celina Heliasz and Maciej Ogrodniczuk. 2019. [Eksplicytność a implicytność w świetle analizy korpusowej \(meta\)tekstu](#). *Linguistica Copernicana*, 16:75–100.
- Jerry R. Hobbs. 1985. On the coherence and structure of discourse. Technical report, CSLI-85-37, Center for the Study of Language and Information.
- Mikel Iruskieta, Arantza Díaz de Ilarraza, and Mikel Lersundi. 2014. [The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 466–475, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- ISO. 2012. ISO 24612. 2012. Language resource management, Linguistic annotation framework. Standard, International Organization for Standardization, Geneva, CH.
- ISO. 2016. ISO 24617-8. 2016. Language resource management, part 8: Semantic relations in discourse (DR-Core). Standard, International Organization for Standardization, Geneva, CH.
- ISO. 2020. ISO 24617-2. 2020. Language resource management-Semantic annotation framework (SemAF) - part 2 - Dialogue acts. Standard, International Organization for Standardization, Geneva, CH.
- Ekaterina Lapshinova-Koltunski and Kerstin Anna Kunz. 2014. Annotating cohesion for multilingual analysis. In *Proceedings of the 10th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 57–64, Reykjavik, Iceland. Association for Computational Linguistics.
- William Mann and Sandra Thompson. 1988. [Rhetorical structure theory: Toward a functional theory of text organization](#). *Text*, 8:243–281.
- William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, Marina del Rey, CA: Information Sciences Institute.
- Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.
- Lester James V. Miranda. 2023. [Developing a named entity recognition dataset for tagalog](#).
- Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawistawska. 2015. [Coreference in Polish: Annotation, Resolution and Evaluation](#). Walter De Gruyter.
- Maciej Ogrodniczuk, Aleksandra Tomaszewska, Daniel Ziembicki, Sebastian Żurowski, Ryszard Tuora, and Aleksandra Zwierzchowska. forthcoming. Polish Discourse Corpus (PDC): Corpus Design, ISO-Compliant Annotation, Data Highlights, and Parser Development. In *Proceedings of The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC COLING 2024)*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Rashmi Prasad and Harry Bunt. 2015. [Semantic relations in discourse: The current state of ISO 24617-8](#). In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. [The Penn Discourse Treebank 2.0](#). In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

- Laurent Prévot, Laure Vieu, and Nicholas Asher. 2009. Une formalisation plus précise pour une annotation moins confuse: la relation d'élaboration d'entité. *J. Fr. Lang. Stud.*, 19(2):207–228.
- Ted Sanders, Wilbert Spooren, and Leo Noordman. 1992. [Toward a taxonomy of coherence relations](#). *Discourse Processes*, 15(1):1–35.
- Deborah Schiffrin. 1987. *Discourse markers*. 5. Cambridge University Press.
- Purificação Silvano, João Cordeiro, António Leal, and Sebastião Pais. 2023. [DRIPPS: a corpus with discourse relations in perfect participial sentences](#). In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 470–481, Vienna, Austria. NOVA CLUNL, Portugal.
- Purificação Silvano and Mariana Damova. 2023. [ISO-DR-core plugs into ISO-dialogue acts for a cross-linguistic taxonomy of discourse markers](#). In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 440–448, Vienna, Austria. NOVA CLUNL, Portugal.
- Purificação Silvano, Mariana Damova, Giedre Valunaite Oleskeviciene, Chaya Liebeskind, Christian Chiarcos, Dimitar Trajanov, Ciprian-Octavian Truica, Elena Simona Apostol, and Anna Bączkowska. 2022. [Iso-based annotated multilingual parallel corpus for discourse markers](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 2739–2749. European Language Resources Association.
- Maite Taboada and Debopam Das. 2013. [Annotation upon annotation: Adding signalling information to a corpus of discourse relations](#). *Dialogue Discourse*, 4:249–281.
- Teun A. van Dijk. 1979. [Pragmatic connectives](#). *Journal of Pragmatics*, 3(5):447–456.
- Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. [Discourse structure and language technology](#). *Natural Language Engineering*, 18(4):437–490.
- Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.
- Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. [Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Sebastian Żurowski, Daniel Ziembicki, Aleksandra Tomaszewska, Maciej Ogrodniczuk, and Agata Drozd. 2023. [Adopting ISO 24617-8 for Discourse Relations Annotation in Polish: Challenges and Future Directions](#). In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 482–492, Vienna, Austria. NOVA CLUNL, Portugal.

8. Language Resource References

- Przepiórkowski, Adam and Bańko, Mirosław and Górski, Rafał L. and Lewandowska-Tomaszczyk, Barbara. 2012. *Narodowy Korpus Języka Polskiego [En. National Corpus of Polish]*. Wydawnictwo Naukowe PWN. PID <http://nkjp.pl/>.

9. Appendix

| | DR-core relations | Definition | Semantic Role | |
|----------------------------|--|---|------------------------|---------------------|
| | | | Arg 1 | Arg2 |
| asymmetric | Cause | Arg2 is an explanation for Arg1. | result | reason |
| | Expansion | Arg2 is a situation involving some entity/entities in Arg1, expanding the narrative of which Arg1 is a part, or expanding on the setting relevant for interpreting Arg1. The Arg1 and Arg2 situations are distinct. | narrative | expander |
| | Asynchrony | Arg1 temporally precedes Arg2. | before | after |
| | Concession | An expected causal relation between Arg1 and \neg Arg2 is cancelled or denied by Arg2. | expectation raiser | expectation-denier |
| | Elaboration | Arg1 and Arg2 are the same situation, but Arg2 provides more detail. | broad | specific |
| | Exemplification | Arg1 is a set of situations; Arg2 is an element of that set. | set | instance |
| | Manner | Arg2 specifies how Arg1 comes about or occurs. | achievement | means |
| | Condition | Arg2 is an unrealized situation which, when realized, would lead to Arg1. | Consequent | Antecedent |
| | Negative Condition | Arg2 is an unrealized situation which, when “not” realized, would lead to Arg1. | Consequent | Negated-Antecedent |
| | Purpose | Arg2 is the goal or purpose of the situation described by Arg1. | Enablement | Goal |
| | Exception | Arg2 indicates one or more circumstances in which the situation(s) described by Arg1 does not hold. | Regular | Exclusion |
| | Substitution | Arg1 and Arg2 are alternatives, with Arg2 being the favored or chosen alternative. | Disfavored-alternative | Favored-alternative |
| | Functional dependence | Arg2 is a dialogue act with a responsive communicative function; Arg1 is the dialogue act(s) that Arg2 responds to. | Antecedent-act | Dependent-act |
| Feedback dependence | Arg2 is a feedback act that provides or elicits information about the understanding or evaluation by one of the dialogue participants of Arg1. | Feedback-scope | Feedback-act | |

Asymmetric discourse relations (ISO, 2016; Bunt and Prasad, 2016).

| | DR-core relations | Definition |
|-----------|--------------------|--|
| symmetric | Conjunction | Arg1 and Arg2 bear the same relation to some situation evoked in the discourse, explicitly or implicitly. Their conjunction indicates that they both hold with respect to that situation. |
| | Contrast | One or more differences between Arg1 and Arg2 are highlighted with respect to what each predicates as a whole or to some entities they mention. |
| | Synchrony | Some degree of temporal overlap exists between Arg1 and Arg2. All forms of overlap are included. |
| | Similarity | One or more similarities between Arg1 and Arg2 are highlighted with respect to what each predicates as a whole or to some entities they mention. |
| | Disjunction | Arg1 and Arg2 bear the same relation to some other situation evoked in the discourse, explicitly or implicitly. Their disjunction indicates that they are alternatives with respect to that situation, with the disjunction being non-exclusive so that both Arg1 and Arg2 may hold. |
| | Restatement | Arg1 and Arg2 describe the same situation, but from different perspectives. |

Symmetric discourse relations (ISO, 2016; Bunt and Prasad, 2016).

Combining semantic annotation schemes through interlinking

Harry Bunt

Department of Cognitive Science and Artificial Intelligence,
School of Humanities and Digital Sciences, Tilburg University
harry.bunt@tilburguniversity.edu

Abstract

This paper explores the possibilities of using combinations of different semantic annotation schemes. This is particularly interesting for annotation schemes developed under the umbrella of the ISO Semantic Annotation Framework (ISO 24617), since these schemes were intended to be complementary, providing ways of indicating different semantic information about the same entities. However, there are certain overlaps between the schemes of SemAF parts, due to overlaps of their semantic domains, which are a potential source of inconsistencies. The paper shows how issues relating to inconsistencies can be addressed at the levels of concrete representation, abstract syntax, and semantic interpretation.

Keywords: semantic annotation, ISO standards, combination of annotation schemes, interlinking

1. Introduction

Existing semantic annotation schemes are nearly always focussed on a specific type of semantic information, such as TimeML (Pustejovsky, 2003) on time and events, SpatialML (Mani et al., 2010) on spatial information, DAMSL (Allen & Core, 1997) on dialogue acts, PDTB (Prasad et al, 2008; 2019) on discourse relations, and RAF (Reference Annotation Framework, Salmon-Alt & Romary, 2005) on coreference. In a similar vein, the ISO Semantic Annotation Framework (ISO 24617, ‘SemAF’) was set up as a multi-part standard, with different parts focussing on different semantic domains. Table 1 lists the SemAF parts that have defined an annotation schema, with an indication of their semantic domain in the leftmost column. The second column specifies the SemAF part number, so for example the part that focuses on the annotation of time and events has defined the standard schema ISO 24617-1, the part for annotating dialogue acts the standard ISO 24617-2, and so on. The third column contains an unofficial name of the standard, which is often used for being mnemonically easier than the official ISO number. The rightmost column indicates some of the most important sources of each SemAF part.

Developing the SemAF standard as a set of separate sub-standards has proved useful, as it is more feasible to develop an annotation schema for a well-delineated semantic domain, and can benefit from the participation of different groups of experts for different domains. The first two parts of SemAF, informally known as ‘ISO-TimeML’ and ‘DiAML’, respectively, are successful examples of the application of this approach, as the annotation of time and events is clearly separable from the annotation of dialogue acts. However, some of the semantic domains are not entirely disjoint.

The annotation schemes of the various SemAF parts are therefore not entirely complementary, and some semantic phenomena are covered in more than one sub-standard. More specifically, semantic phenomena that play central stage in one domain may play a peripheral role in another domain. For example, the temporal expression “*every Monday*” quantifies over Mondays. Being a temporal expression, ISO-TimeML provides an annotation of this expression, including an indication of its quantifying character. ISO-TimeML has only a rudimentary treatment of quantification, however (Bunt & Pustejovsky, 2016), while it is the focus of SemAF part 12, QuantML.¹

The marginal treatment of temporal quantification can be seen as a limitation of ISO-TimeML; on the other hand, ISO-TimeML offers a more detailed treatment of events and temporal entities than QuantML, which can be seen as a limitation of QuantML. Limitations of this kind are no problem when annotating language data with (a) information about events and time, or (b) about quantifications, but they present a problem for annotating data about *both* quantifications *and* time and events. In the latter case, one would like to combine the possibilities offered by the two annotation schemes. One way to do this is to define a new annotation schema that makes use of elements from the two schemes. In this paper we explore another idea: the combination of annotations provided by two (or more) annotation schemes without modifying them, but adding links between elements of the annotations in order to express that the two schemes annotate the same primary data with a different focus.

¹At the time of writing, QuantML was the subject of a ballot for obtaining the status of an international ISO standard. See also Bunt (2024).

| Semantic domain | # | Name | Source |
|-------------------------------------|-----|------------|--|
| Time and Events | 1 | ISO-TimeML | TimeML (Pustejovsky, 2003) |
| Dialogue acts | 2 | DiAML | DIT++ (Bunt, 2007) |
| Semantic roles | 4 | ISO-SR | LIRICS and VerbNet, (Palmer & Bunt 2013, Bonial et al. 2011) |
| Spatial information | 7 | SpaceML | SpatialML (Mani et al., 2010; Pustejovsky & Lee, 2015) |
| Dscourse relations | 8 | DR-Core | PDTB (Prasad et al, 2008, 2019) |
| Coreference | 9 | ISO-RAF | RAF, Reference Annotation Framework (Salmon-Alt & Romary, 2005) |
| Measurable Quantitative Information | 11 | MQI | (Hao et al., 2019) |
| Quantification | 12: | QuantML | (Bunt, 2019a) (under review) |

Table 1: SemAF parts that have defined an annotation schema

The idea of this technique, ‘*interlinking*’, is very simple: given two annotation schemes A and B which represent different information about the same event or other kind of entity, interlinking adds to the A- and B-annotations an identity relation between the corresponding elements. This is illustrated in Figure 2, where a mini-discourse is annotated with TimeML, QuantML, and DR-Core, which all use XML-based representations, with <idLink>s indicating that the same three events are annotated in each of the three schemes.

This paper is organised as follows. Section 2 discusses related work. Section 3 summarises the ISO Semantic Annotation Framework as far as relevant for the present study, and explores overlaps and inconsistencies between SemAF parts. Section 4 specifies the mechanism of *interlinking*, with detailed examples. Section 5 summarises the present study, including its limitations, and an outlook of future work.

2. Related Work

The interest in combining annotation schemes has three main reasons.

First, specialised annotation schemes restricted to a specific semantic domain, like those of the SemAF parts, has the danger of designing schemes that have certain gaps, which may limit the coverage of individual annotation schemes in unwelcome ways for corpus annotation. Examples of such gaps are:

- (1) anaphorically expressed participants in events cannot be annotated in QuantML, ISO-TimeML, and SpaceML (other than by simply assuming anaphora to have been resolved);
- (2) temporal and spatial quantification have no adequate treatment in ISO-TimeML and SpaceML (Bunt & Pustejovsky, 2016);
- (3) although semantic roles play a central role in QuantML annotations, they are undefined there - that is the subject matter of ISO-SR.

Some of these gaps could be resolved by combining SemAF annotation schemes, such as ISO-TimeML and QuantML, or SpaceML and ISO-RAF, or QuantML and ISO-SR.

Second, semantic annotation may play an important role in applications which require not just the annotation of one semantic domain, such as time and events, but also of other domains, such as coreference and discourse relations. This is for example the case in an application discussed by Silvano (2021) and Leal (2022), who used concepts from different SemAF annotation schemes to design a new, integrated schema to meet the requirements of the application. The design of integrated annotation schemes is also addressed in Malchanau et al. (2024).

Third, the markup language of an annotation schema may be used not only for the annotation of corpus data, but also as an internal interface language in an NLP system. For example, the dialogue act markup language DiAML has been used as an internal language in which the modules of an interactive language-based system communicate, in particular as an interface language for dialogue management (Malchanau, 2019). When used for this purpose, a notable limitation of DiAML is that, while it supports a rich annotation of dialogue acts, their communicative functions, and relations between them, it does not provide a way to indicate their semantic content. This limitation has been addressed by Bunt (2019), who proposed the use of *annotating schema plug-ins* for adding descriptive (and semantic) power to a host annotation schema.

Besides the definition of integrated schemes that combine elements from different schemes, which and the addition of plug-ins to a host annotation schema, another option is explored in this paper, in which existing annotation schemes are used in combination without altering them,.

3. The Semantic Annotation Framework

3.1. Architecture of SemAF Parts

All parts of SemAF follow the same architecture, described in ISO 24617-6: Principles of semantic annotation see also Bunt (2015) and Pustejovsky et al. (2017). QuantML thus has a triple-layered definition consisting of:

1. An abstract syntax, which specifies the class of well-defined *annotation structures* as pairs, triples, and other set-theoretical constructs containing quantification-related concepts. Annotation structures consist of *entity structures*, which contain information about a stretch of primary data, and *link structures*, which contain information relating two (or more) entity structures. The role of the abstract syntax is visualized in Figure 1.
2. A semantics, which specifies the meaning of the annotation structures defined by the abstract syntax. QuantML has an interpretation-by-translation semantics, which translates annotation structures to discourse representation structures (DRSs, Kamp & Reyle, 1993). The use of DRSs is mainly motivated by the fact that this formalism is also used in other SemAF parts.
3. A concrete syntax, that specifies a representation format for annotation structures. The QuantML definition includes an XML-based reference format, again mainly motivated by the use of XML in other standards.

The three levels are interrelated by encoding (F_{ac}), decoding (F_{CA}), and interpretation functions; see Figure 1. Since the semantics is defined at the level of the abstract syntax, alternative representation formats may be used that share the same abstract syntax, as indicated in Figure 1 and are thus semantically equivalent. This adds to the interoperability of the schema.

3.2. Complementarity of SemAF parts

The various parts of SemAF are intended to be complementary, dealing with different semantic domains. However, as noted above, these domains often have overlaps, which is a potential source of inconsistencies. In particular, because of the common event-based semantic approach, events and their participants and the relations between them play a role in several SemAF parts. The following example highlights some of these overlaps, showing the information that six SemAF parts would annotate for the mini-discourse of (1a).

- (1) a. After moving the pianos to the stage, the men had a beer. They were thirsty.

- b. **ISO-TimeML:** a **move event** occurred, followed by a **beer-drinking event** which occurred **in the past**. A **be-thirsty event** occurred **in the past**.

ISO-SR: a **move event** occurred with pianos as *Themes* and a stage as *Final Location*. A **drinking event** occurred with some men as *Agent(s)* and some beer as *Patient*. A **be-thirsty event** occurred, with certain individuals as *Experiencers*.

SpaceML: a **move event** occurred with a stage as **end point**.

DR-Core: a **move event** occurred which *caused* a **be-thirsty event**, which *explains* the occurrence of a **beer-drinking event**.

ISO-RAF: the set of *discourse entities* that “they” refers to *is the same as* the *it set referred to* by “the men”.

QuantML: some **move events** occurred in which certain *contextually determined men* participated *collectively* as an **Agent**. The *men* acted *individually* as the *Agent* in **drinking events** with *some beer* as **Patient**. A **be-thirsty event** occurred, with *certain individuals* as **Experiencers**.

This example clearly shows that each of the annotation schemes focuses on different information, but information concerning events with their participants and relations plays a role in nearly all of them. In the next subsection we consider the consequences of these overlaps.

3.3. Overlaps of SemAF parts

3.3.1. Events

Events play central stage in ISO-TimeML, in which they have articulate annotations as illustrated in example (3). Events that involve motion are equally important in SpaceML, and have a similar articulate annotation there. For annotating events expressed by verbs, ISO-SR makes use of ‘eventuality frames’, borrowed from VerbNet, which allows distinctions to be made between different verb senses. ISO-TimeML proposes articulate annotations both for events described by verbs and for events described by nouns. QuantML and DR-Core treat events, regardless of their lexical description, as predicate constants (in the spirit of DRT and other formal semantic approaches).

Example (2) shows annotations of a *call* event in the sentence *Peter called this morning* represented in each of these annotation schemes. The value

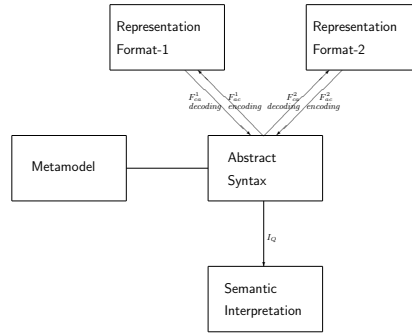


Figure 1: Architecture of SemAF parts.

(2) a. Peter called this morning.

b. **Representation of events** in various SemAF parts:

ISO-TimeML: `<event xml:id="e1" target="#w2" pred="call" class="occurrence" type="transition" pos="verb" tense="present" aspect="perfective" mood="none" polarity="positive" modality="certain"/>`

SpaceML: `<event>` as in ISO-TimeML, with additional attributes (`@latLong`, `@elevation`,...)

ISO-SR: `<eventuality xml:id="e1" target="#m2" eventFrame="#call.03"/>`

DR-Core, MQI: `<event xml:id="e1" target="#m2" type="call"/>`

QuantML: `<event xml:id="e1" target="#m2" pred="call" repetitiveness="1"/>`

'call.03' of the `@eventFrame` attribute in the ISO-SR annotation is assumed to identify the event frame for the intended sense of *call*, i.e. referring to an event that could also be described by the verb *to phone*.

To what extent are these alternative representations consistent? An important point to note is that all 6 annotations represent the *same* event, expressed in the primary data by the markable 'm2'. The ISO-TimeML representation just adds more information about the type of event and the way it is described in the primary data. A semantic difference between the ISO-TimeML and QuantML representations might seem to be that the latter is interpreted as a set of one or more events, whereas the ISO-TimeML representation refers to a single event. This is not quite the case, however, since the semantics of ISO-TimeML is defined by means of an existential quantifier, saying that *there has been a call-event such that...*, without ruling out that more than one event of the same type occurred. In this respect the two representations are therefore semantically equivalent. The additional `@repetitiveness` attribute in QuantML is used to accommodate expressions like *called twice*, indicating the cardinality of a set of events. If an annotation is intended to indicate the occurrence of a single event, this can

be expressed in QuantML by the `@repetitiveness` attribute having the value '1'.

The fact that the various annotations represent the same concept, though possibly with more or less detail, will be essential for the interlinking mechanism described in the next section.

3.3.2. Participants

The entities that participate in events can be divided into (1) temporal and spatial entities, (2) events, (3) (measurable) quantities, and (4) objects of any other kind. Events participating in other events have the same articulate representation as the events in which they participate. Non-eventive entities have an articulate annotation in ISO-RAF, as shown in example (3). Entities of any kind (temporal, spatial, eventive, quantitative, other) occurring as participants in events all have articulate representations in QuantML; see example (3).

QuantML annotates the distinction between collective and individual (or 'distributive') quantification which is illustrated in example (1) if we assume that *the men* collectively moved *the pianos* and individually had a beer; therefore, participants in QuantML are represented by `<entity>` elements interpreted as sets.

(3) a. Peter called this morning.

b. **Representation of entities** as participants in events or inter-entity relations:

ISO-TimeML :

```
<timex3 xml:id="x1" target="#m3" pred="morning"... />
```

ISO-RAF :

```
<discourseEntity xml:id="e1" target="#m1" abstractness="concrete"
referentialStatus="discourseNew" animacy="animate" ... />
<discourseEntity xml:id="e1" target="#m3" animacy="inanimate" ... />
```

QuantML :

```
<entity xml:id="x1" target="#m1" involvement="all"
individuation="count" size="1"/>
<refDomain xml:id="x2" target="#m1" pred="peter" determinacy="det"/>
<entity xml:id="x3" target="#m3" involvement="all"
individuation="count" size="1"/>
<refDomain xml:id="x4" target="#m3" pred="morning"
determinacy="det"/>
```

c. **Representation of relations** between events, participants, and time, as annotated above and in (2):

ISO-TimeML :

```
<tLink eventID="#e1" relatedToTime="#x3" relType="isIncluded"/>
```

QuantML :

```
<participation event="#e1" participant="#x1" semRole="aevent"/>
<participation event="#e1" participant="#x3" semRole="time"/>
```

Example (3) shows annotations of the participants in example (1). ISO-TimeML only provides a representation for the temporal expression *this morning*; ISO-RAF and QuantML provide a representation for both *Peter* and *this morning*. The QuantML representation indicates that both NPs are countable (as opposed to the mass NP *some beer* in example (1)), that both NPs quantify over a definite domain, consisting of only one individual in the case of the NP *Peter*, and that all the members of both domains participate in the event(s) under discussion.

3.3.3. Relations

The following SemAF parts annotate relations among events, participants, time and place:

ISO-TimeML represents (1) information about the time of occurrence of events; (2) temporal relations between events, as expressed by conjunctions of clauses or by a main clause and a subordinate clause; (3) temporal relations between temporal objects. All these relations are represented using <tLink> elements.

SpaceML represents (1) spatial information about the occurrence of events, including locations of begin and end points, trajectories and paths of movements, (2) spatial relations between spatial objects, using a variety of links.

ISO-SR represents relations between events and participants in terms of semantic roles.

QuantML uses the semantic roles defined in ISO-SR as attribute values in <participation> links, and moreover represents (1) non-temporal semantic relations between events, as expressed by a main clause and a subordinate clause; (2) relations between any two kinds of entities as expressed by noun-noun modifiers, possessives, prepositional phrases, or relative clauses, using various links, such as <nnMod>, <ppMod>, and <possMod>.

DR-Core represents semantic relations such as *Cause*, *Contrast*, *Concession*, *Elaboration* between events as expressed in a discourse by clauses either within the same sentence or in different sentences.

Inspecting the information represented in these annotation schemes, we can again see a great deal of complementarity, but also some overlaps, and hence a danger of inconsistencies. We discuss these in the next subsection.

3.4. Levels of inconsistency

The various SemAF parts display inconsistencies in representing the same information in different ways, or as representing more detailed and different information about the same events, entities, or relations. To what extent do the inconsistencies

“After moving the pianos to the stage, the men had a beer. They were thirsty.”

Markables: m1 = “After”, m2 = “moving”, m3 = “the piano”, m4 = “to”, m5 = “the stage”,
m6 = “the men”, m7 = “had” m8 = “a beer”, m9 = “They”, m10 = “were”,
m11 = “were thirsty” m12 = ”thirsty”

QuantML:

```
<entity xml:id="xQ1" target="#m3" refDomain="#xQ2" individuation="count"
  involvement="all"/>
<refDomain xml:id="xQ2" target="#m3" pred="piano" determinacy="det"/>
<entity xml:id="xQ3" target="#m5" refDomain="#xQ4" individuation="count" size="1"
  involvement="all"/>
<refDomain xml:id="xQ4" target="#m5" pred="stage" determinacy="det"/>
<event xml:id="eQ1" target="#m2" pred="move"/>
<participation event="#eQ1" participant="#xQ1" semRole="theme"/>
<participation event="#eQ1" participant="#xQ3" semRole="finalLocation"/>
<entity xml:id="xQ5" target="#m6" refDomain="#xQ6" individuation="count" .../>
<refDomain target="#m6" pred="man" determinacy="det"/ ... />
<participation event="#eQ2" participant="#xQ5" semRole="agent"/>

<event xml:id="eQ2" target="#m7" pred="drink"/>
<entity xml:id="xQ7" target="#m8" refDomain="#xQ8" individuation="count"
  involvement="some"/>
<refDomain target="#m8" pred="beer" determinacy="indet"/>
<participation event="#eQ2" participant="#xQ5" semRole="patient"/>
<event xml:id="eQ3" target="#m10" pred="be"/>
<predication event="#eQ3" participant="#xQ1" predicate="thirsty" distr="individual"/>
```

ISO-TimeML:

```
<event xml:id="eT1" target="#m2" pred="move" .../>
<event xml:id="eT2" ptarget="#m7" pred="drink" ... tense="past" />
<event xml:id="eT3" ptarget="#m10" pred="be-thirsty" .../>
<signal xml:id="s1" target="#m1" pred="after"/>
<tLink arg1="#eT1" arg2="#eT1" relType="after"/>
```

DR-Core:

```
<event xml:id="eD1" target="#m2" pred="move" .../>
<event xml:id="eD2" target="#m7" pred="drink" ... tense="past" ... />
<event xml:id="eD3" target="#m10" pred="be-thirsty" ... tense="past" ... />
<drLink arg1="#eD2" arg2="#eD1" relType="succession"/>
<drLink arg1="#eD3" arg2="#eD2" relType="cause"/>
```

Interlinking ISO-TimeML to QuantML:

```
<idLink arg1="#eQ1" arg2="#eT1"/>
<idLink arg1="#eQ2" arg2="#eT2"/>
<idLink arg1="#eQ3" arg2="#eT3"/>
```

Interlinking DR-Core to ISO-TimeML:

```
<idLink arg1="#eD1" arg2="#eT1"/>
<idLink arg1="#eD2" arg2="#eT2"/>
<idLink arg1="#eD3" arg2="#eT3"/>
```

Figure 2: Example of interlinking at the level of concrete syntax.

noted above actually present a problem? So far, we discussed inconsistencies at the level of concrete (XML-based) representation; the addition of interlinking `<idLink>` elements (or a similar device in other representation formats) seems relatively straightforward, and the intuitive meaning of the interlinks is simple and clear, but they might cause inconsistencies at the deeper levels of abstract syntax and semantics. To remain in line with the ISO principles of semantic annotation

(ISO 24617-6), the entire structure formed by the concatenation of the representations of interlinked schemes and the links between them should have a well-defined abstract syntax with a semantic interpretation.

The inconsistencies between SemAF parts, due to overlapping semantic domains, can be divided into three categories:

1. Different terms used for the same concept, e.g. the attribute @pred in some of the schemes is called @type in others.
2. Different sets of attributes and values used to describe the same events or other entities, reflecting the focus of different schemes.
3. Different views on how events and other entities are conceptually related.

Inconsistencies of type (1) arise purely at the level of concrete syntax, have no semantic consequences, and may be considered trivial. The decoding function that computes the abstract syntax of interlinked annotations can simply map equivalent terms to the same concepts in the abstract syntax. Inconsistencies of type (2) are potentially more serious, but not necessarily so. They are not problematic if the differences in sets of attributes correspond to semantically complementary information, or if one set of attributes and values is semantically more specific than another. An interesting case is the difference between ISO-TimeML and SpaceML on the one hand, and ISO-SR and QuantML on the other, regarding the annotation of relations between events and their time and place of occurrence. ISO-SR includes 4 temporal relations: *Time*, *Initial-Time*, *Final-Time*, and *Duration* and 5 spatial relations: *Location*, *Initial-Location*, *Final-Location*, *Distance*, *Path*, ISO-TimeML, by contrast, makes use of 7 relations: *Simultaneous*, *Includes*, *IsIncluded*, *Before*, *I-Before*, *After*, *I-After* (where I-Before and I-After mean immediately before and immediately after, respectively, and SpaceML has a large set of spatial relations. These differences reflect that ISO-TimeML and SpaceML have the domains of time and space as their respective focus, and these are semantically not problematic, since the relations of ISO-SR are less specific than those of ISO-TimeML and SpaceML, so the former entail the latter. This makes the ‘inconsistency’ semantically harmless (although somewhat redundant).

Inconsistencies of type (3) are the most fundamental, and are often the cause of a type (2) inconsistency. This is for example the case for temporal relations among events and for relations between events and time of occurrence. These cases, and all other cases in SemAF that we have examined, can all be treated in the same way as type (2) inconsistencies. Example (7) shows that interlinking can be used to accommodate different conceptual views at the level of concrete representations while providing a consistent semantic interpretation.

4. Interlinking

4.1. Concrete syntax

The example in Figure 3.3.3. illustrates the use of interlinking for the annotation structure that com-

bines elements from ISO-TimeML and QuantML, where a mini-discourse is annotated with TimeML, QuantML, and DR-Core, with <idLink>s indicating that the same events are annotated in all three schemes.

4.2. Abstract Syntax

The decoding function of an annotation schema, which computes the abstract syntax of the concrete representation (see Fig. 2) uses the interlinking specifications to merge the semantic information about the same events and the same entities that occur in the respective annotations.

In QuantML, the unit of annotation is a clause. At the abstract syntax level, a clause annotation structure is a quadruple of the form (4), consisting of specifications of (1) an event; (2) a set of n participants ($n > 0$) (3) a set of n participation links; and (4) a set of $n - 1$ scope links.

$$(4) A_Q = \langle \epsilon_e, \{\epsilon_1, \dots, \epsilon_n\}, \{L_1, \dots, L_n\}, \{s_1, \dots, s_{n-1}\} \rangle.$$

The abstract syntax of the annotations of other SemAF-parts that annotate events and participating entities are the same as (4) for a simple clause, except that the set of scope links is empty, as they do not annotate scope relations. Moreover, ISO-TimeML and SpaceML consider only temporal and spatial entities, and hence use specific time- and space-related relations rather than general participation relations. The interlinking of two or more of these annotation schemes has the effect of creating another annotation structure in the general quadruple form of (4), as follows.

Let X_A and X_B be the XML-representations of a clause, annotated according to the annotation schemes A and B, and X_{IL} the set of statements that interlink X_A and X_B . Application of the decoding functions F_{ca}^A and F_{ca}^B can be represented schematically as follows:

$$(5) \begin{aligned} F_{ca}^A(X_A) &= \langle \epsilon_A, E_A, L_A, sc_A \rangle, \\ F_{ca}^B(X_B) &= \langle \epsilon_B, E_B, L_B, sc_B \rangle \end{aligned}$$

Let \mathcal{R}_{IL} be the function that replaces in a given set of expression sll occurrences of an identifier x_i which occurs either as first or as second item in the set of pairs $F_{ca}^{AB}(X_{IL})$ by the corresponding pair $\langle \epsilon_{xAi}, \epsilon_{xBj} \rangle$ (where $\epsilon_{xAi} \in E_A$ and $\epsilon_{xBj} \in E_B$).

The decoding function F_{ca}^{AB} of the interlinked schemes constructs pairs of elements that correspond to the arguments of an <idLink> in the concrete syntax applied to the set X_{IL} of interlinks, a set of corresponding pairs $\langle \epsilon_{xAi}, \epsilon_{xBj} \rangle$ is constructed, where $\epsilon_{xAi} \in E_A$ and $\epsilon_{xBj} \in E_B$.

Using ‘+’ to indicate concatenation, the decoding function applied to the entire XML representation $X_A + X_B + X_{IL}$, is defined as:

- (6) $F_{ca}^{AB}(X_A + X_B + X_{IL}) = \langle \epsilon_{AB}, E_{AB}, L_{AB}, sc_{AB} \rangle$, with
- $\epsilon_{AB} = \langle \epsilon_A, \epsilon_B \rangle$,
 - $E_{AB} = \mathcal{R}_{IL}(E_A) \cup \mathcal{R}_{IL}(E_B)$,
 - $L_{AB} = \mathcal{R}_{IL}(L_A) \cup \mathcal{R}_{IL}(L_B)$,
 - $sc_{AB} = \mathcal{R}_{IL}(sc_A) \cup \mathcal{R}_{IL}(sc_B)$

The set L_{AB} of event - entity links and the set of scope links sc_{AB} are computed in the same way as the set of entities E_{AB} , by merging the corresponding components of the linked schemes after replacing single entities by pairs in case they are interlinked.

4.3. Semantics

The semantic interpretation of interlinked A - and B -annotations is computed by the interpretation function I_{AB} , defined in terms of the interpretation functions I_A and I_B . Central in the definition of I_{AB} is the interpretation of pairs of events or pairs of participants which were linked by $\langle idLink \rangle$ s in the XML representation and which occur as participant pairs in the abstract syntax, simply as the merge of the two interpretations.²

$$(8) I_{AB}(\epsilon_A, \epsilon_B) = I_A(\epsilon_A) \cup I_B(\epsilon_B)$$

The semantic interpretation of a fully connected annotation schema, in which the relative scopes of all participants are specified, can be computed by combining the interpretations of all the event - entity link structures, since these structures embed the event structures and entity structures that describe the events and participants. This can be done in a compositional manner, using the semantics of scope links to determine how the interpretations of event and entity structures are combined; this has been worked out in detail for the semantics of QuantML (Bunt, 2023). The upshot of this is expressed in (9), where the set L_{AB} of link structures is ordered by their relative scopes; σ_{ij} is the composition function that is computed by applying I_{AB} to the corresponding scope relation in the abstract syntax.

$$(9) I_{AB}(\epsilon_{AB}, E_{AB}, L_{AB}, sc_{AB}) = I_{AB}(L_{AB}) = \\ = I_{AB}(L_1, L_2, \dots, L_n) \\ = \sigma_{12}(I_{AB}(L_1), \sigma_{23}(I_{AB}(L_2), \dots \\ I_{AB}(\sigma_{n-1,n}(I_{AB}(L_n)) \dots))$$

Example (7) shows in detail how this works out for the sentence *Ninety-five students graduated on a Friday*, instantiating the ‘A’ and ‘B’ in

²This notation assumes interpretations to have the form of DRSs. ISO-TimeML has a semantics defined in different terms, which is however readily converted to DRS form.

(5), (8), and (9) by ‘Q’ (for QuantML) and ‘T’ (for ISO-TimeML). The abstract syntax of the XML representation, computed by the decoding function F_{ca}^{QT} , is shown in (7b); its semantics as calculated by the interpretation function I_{QT} is shown in (7c) (where \cup^* is a scope-preserving merge operation on DRSs; see Bunt, 2023). The XML representations, are slightly simplified to save space.

The final semantic interpretation, formulated as the DRS in (10), effectively says that there is a set (‘X’) of 95 students for whom there is a set of 1 friday, for which the description “XXXX-WXX-5” applies, which have graduation events as their time of occurrence, and include the time of occurrence. This combines the information in the QuantML and ISO-TimeML annotations. There is some redundancy in the final result, but such semantic redundancy is perhaps not very elegant, but formally harmless.

$$(10) [X \mid |X|=95, x \in X \rightarrow [\text{student}(x), \\ [Y \mid |Y|=1, y \in Y \rightarrow [\text{friday}(y), \\ \text{value}(y) = \text{“XXXX-WXX-5”}, \\ [E \mid e \in E \rightarrow [\text{graduate}(e), \\ \text{class}(e) = \text{occurrence}, \text{type}(e) = \text{transition}, \\ \text{agent}(e,x), \text{time}(e,y), \text{is_included}(e,y)]]]]]]]$$

5. Conclusion and Further Work

In this paper we have presented an exploration of the possibilities of using combinations of semantic annotation schemes. This seems particularly interesting for the use of annotation schemes developed under the umbrella of the ISO Semantic Annotation Framework, since these schemes were intended to be complementary, serving to express information in different semantic domains. The schemes developed as SemAF parts have certain unavoidable overlaps, however, due to unavoidable overlaps of semantic domains, which are a source of potentially problematic inconsistencies and which may be harmful for their interoperability.

For truly complementary schemes, like DiAML, QuantML, and DR-Core, the interlinking technique seems perfectly suitable. For interlinking annotations of overlapping schemes, such as ISO-TimeML and QuantML, we have shown promising possibilities for constructing semantically consistent interlinked annotations, but a more elaborate exploration of all the overlaps in SemAF parts is needed to fully evaluate this proposal.

(7) “Ninety-five students graduated on a Friday”

Markables: m1 = “Ninety-five”, m2 = “Ninety-five students”, m3= “students”, m4= “graduated ”, m5 = “on”, m6 = “on a Friday”, m7 = “a Friday”, m8 = “Friday”

a. XML REPRESENTATION:

X_QuantML

```
<entity xml:id="xQ1" target="#m2" refDomain="" #x1 involvement="95"
  individuation="count"/>
  <refDomain xml:id="x1" target="#m3" pred="student" determinacy="indet"/>
  <event xml:id="eQ1" target="#m3" pred="graduate"...>
  <participation event="#eQ1" participant="#xQ1" semRole="agent" >
  <entity xml:id="xQ2" target="#m7" refDomain="" #x2 individuation="count"
  involvement="some"/>
  <refDomain xml:id="x2" target="#m3" pred="friday" determinacy="indet"/>
  <participation event="#eQ1" participant="#xQ2" semRole="time"/>
  <scoping arg1="#xQ1" arg2="#xQ2" scopeRel="wider"/>
```

X_ISO-TimetML

```
<event xml:id="eT1" target="#m23 .... pred="graduate" ...>
  <signal xml:id="s1" target="#m5" pred="on"/>
  <timex3 xml:id="xT1" target="#m8" pred="friday" type="date" value="XXXX-WXX-5"/>
  <tLink signalID="#s1" eventID="#eT1" relatedToTime="#xT1" relType="isIncluded"/>
```

X_Interlinking:

```
<idLink arg1="#eQ1" arg2="#eT1"/>
  <idLink arg1="#xQ2" arg2="#eT1"/>
```

b. ABSTRACT SYNTAX:

QuantML:

$$A_Q = \langle eQ1, \{xQ1, xQ2\}, \{pL1, pL2\}, \{pL1, pL2, wider\} \rangle$$

ISO-TimetML:

$$A_T = \langle eT1, \{xT1\}, \{tL1\}, \{\} \rangle$$

Interlinked structure :

$$A_{QT} = \langle \langle eQ1, eT1 \rangle, \{xQ1, \langle xQ2, xT1 \rangle\}, \{pL1, pL2, tL1\}, \{pL1, pL2, wider\} \rangle$$

c. SEMANTICS:

QuantML:

$$\begin{aligned} I_Q(A_Q) &= I_Q(\langle eQ1, \langle xQ1, xQ2 \rangle, \langle pL1, pL2 \rangle, \langle pL1, pL2, wider \rangle \rangle) \\ &= \cup^* (I_Q(xQ1), I_Q(eQ1), \langle I_Q(\text{agent}, \text{individual}) \rangle) \\ &= [X \subseteq \text{student} \mid |X|=95], I_Q([Y \subseteq \text{friday} \mid |Y|=1], I_Q(\text{agent}, \text{individual})) \\ &= [X \subseteq \text{student} \mid |X|=95, x \in X \rightarrow \\ &\quad [Y \subseteq \text{friday} \mid |Y|=1, y \in Y \rightarrow \\ &\quad [E \subseteq \text{graduate} \mid e \in E \rightarrow [\text{agent}(e,x), \text{time}(e,y)]]]] \end{aligned}$$

ISO-TimetML:

$$\begin{aligned} I_T(A_T) &= [Y \subseteq \text{friday} \mid y \in Y \rightarrow [\text{value}(y) = \text{"XXXX-WXX-5"}], \\ &\quad E \subseteq \text{graduate} \mid e \in E \rightarrow [\text{class}(e) = \text{occurrence}, \\ &\quad \text{type}(e) = \text{transition}, \text{is_included}(e,y)]] \end{aligned}$$

Interlinked interpretation:

$$\begin{aligned} I(A_{QT}) &= I_{QT}(\langle \langle pL1, pL2, wider \rangle, \langle pL2, tL1, equal \rangle \rangle) \\ &= \cup^* (I_{QT}(pL1), I_{QT}(pL2, tL1, equal)) \\ &= \cup^* (I_{QT}(pL1), (I_{QT}(pL2) \cup I_{QT}(tL1))) \\ &= \cup^* (I_Q(pL1), (I_Q(pL2) \cup I_T(tL1))) \\ &= [X \mid |X|=95, x \in X \rightarrow [\text{student}(x), [Y \mid |Y|=1, y \in Y \rightarrow [\text{friday}(y), y] = \text{"XXXX-WXX-5"}], \\ &\quad [E \mid e \in E \rightarrow [\text{graduate}(e), \text{class}(e) = \text{occurrence}, \text{type}(e) = \text{transition}, \\ &\quad \text{agent}(e,x), \text{time}(e,y), \text{is_included}(e,y)]]]] \end{aligned}$$

6. Bibliographical References

References

- J. Allen and M. Core. 1997. *DAMSL: Dialogue Act Markup in Several Layers (Draft 2.1)*. Technical Report. University of Rochester, Rochester, NY.
- C. Bonial, W. Corvey, M. Palmer, V. Petukhova, and H. Bunt. 2011. A hierarchical unification of LIRICS and VerbNet semantic roles. In *Proceedings IEEE-ICSC 2011 Workshop on Semantic Annotation for Computational Linguistic Resources*, Stanford, CA.
- H. Bunt. 2009. The DIT++ taxonomy for functional dialogue markup. In *Proceedings of AAMAS 2009 Workshop 'Towards a Standard Markup Language for Embodied Dialogue Acts'*, pages 13–24, Budapest.
- H. Bunt. 2015. On the principles of semantic annotation. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, pages 1–13, London.
- H. Bunt. 2019a. An annotation scheme for quantification. In *Proceedings 14th International Conference on Computational Semantics (IWCS 2019)*, pages 31–42, Gothenburg, Sweden.
- H. Bunt. 2019b. Plug-ins for content annotation of dialogue acts. In *Proceedings 15th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-15)*, pages 34–45, Gothenburg, Sweden.
- H. Bunt. 2023. The Compositional Semantics of QuantML. In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-19)*, pages 56 – 65, Nancy, France.
- H. Bunt. 2024. QuantML: A proposed new standard for semantic annotation. In *Proceedings LREC-COLING 2024, Turin, Italy*, Paris. ELRA.
- D. Davidson. 1967. The Logical Form of Action Sentences. In N. Resher, editor, *The Logic of Decision and Action*, pages 81–95. The University of Pittsburgh Press, Pittsburgh.
- T. Hao, S. Liu, H. Wang, C. Xinyu, and K. Lee. 2019. The semantic annotation of measurable quantitative information. In *Proceedings of the 15th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-15)*.
- ISO. 2012. *ISO 24617-1: 2012, Language Resource Management - Semantic Annotation Framework (SemAF) - Part 1: Time and events*. International Organisation for Standardisation ISO, Geneva.
- ISO. 2014. *ISO 24617-4: 2014, Language Resource Management - Semantic Annotation Framework (SemAF) - Part 4: Semantic roles*. Geneva: International Organisation for Standardisation ISO.
- ISO. 2015. *ISO 24617-6:2015, Language Resource Management - Semantic Annotation Framework (SemAF) - Part 6: Principles of semantic annotation*. International Organisation for Standardisation ISO, Geneva.
- ISO. 2019. *ISO/WD 24617-12:2019, Language Resource Management: Semantic Annotation Framework (SemAF) - Part 12: Quantification*. International Standard. International Organisation for Standardisation ISO, Geneva.
- ISO. 2024. *ISO/WD 24617-12: 2021, Language Resource Management - Semantic Annotation Framework (SemAF) - Part 12: Quantification*. International Organisation for Standardisation ISO, Geneva.
- H. Kamp and U. Reyle. 1993. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht.
- A. Leal, P. Silvano, E. Amorim, I. Cantante and F. Silva, A. Jorg, and R. Campos. 2022. The place of ISO-space in txet2story multilayer annotation scheme. In *Proceedings of the 18th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-17)*, pages 57–66.
- A. Malchanau. 2019. *Cognitive architecture of multimodal multidimensional dialogue management*. Saarbruecken, publisher=Ph.D. dissertation, Saarland University.
- A. Malchanau, V. Petukhova, and H. Bunt. 2024. On the principles of semantic annotation. In *Proceedings 20th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-20)*, Turin, Italy.
- I. Mani, C. Doran, D. Harris, J. Hizeman and R. Quimby, and J. Richer. 2010. SpatialML: annotatopn scheme, resources, and evaluation. *Language Resources and Evaluation*, 44 (3):263–280.
- T. Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press, Cambridge, MA.
- V. Petukhova and H. Bunt. 2008. LIRICS semantic role annotation: design and evaluation of a set of data categories. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. ELRA, Paris.

- R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.
- R. Prasad, B. Webber, A. Lee, and A. Joshi. 2019. *Penn Discourse Treebank Version 3.0*. Linguistic Data Consortium.
- J. Pustejovsky, H. Bunt, and K. Lee. 2010. ISO-TimeML. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta. ELDA, Paris.
- J. Pustejovsky, H. Bunt, and A. Zaenen. 2017. Designing annotation schemes: From theory to model. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 21–72. Springer, Berlin.
- J. Pustejovsky, J. Castano, R. Ingria, R. Gaizauskas, G. Katz, R. Saurí, and A. Setzer. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 337–353, Tilburg, Netherlands.
- S. Salmon-Alt and L. Romary. 2005. The Reference Annotation Framework: A Case for Semantic Content Representation. In *Proceedings of the Sixth International Workshop on Computational Semantics (IWCS-6)*, pages 259–270, Tilburg. Tilburg University Computational Linguistics and AI Group.
- P. Silvano, A. Leal, F. Silva, I. Cantante, F. Oliveira, and A. Jorge. 2021. Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus. In *Proceedings of the 17th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-17)*, pages 1–13.

Fusing ISO 24617-2 Dialogue Acts and Application-Specific Semantic Content Annotations

Andrei Malchanau¹, Volha Petukhova¹, Harry Bunt²

¹ Spoken Language Group, Saarland University, Germany

³ Tilburg Center for Cognition and Communication, Tilburg University, The Netherlands

{andrei.malchanau; v.petukhova}@lsv.uni-saarland.de

harry.bunt@tilburguniversity.edu

Abstract

Accurately annotated data determines whether a modern high-performing AI/ML model will present a suitable solution to a complex dialogue application challenge, without wasting time and resources. The more adequate the structure of incoming data is specified, the more efficient the data can be interpreted and used by the application. This paper presents an approach to an application-specific dialogue semantics design which integrates the dialogue act annotation standard ISO 24617-2 and various domain-specific semantic annotations. The proposed multi-scheme design offers a plausible and a rather powerful strategy to integrate, validate, extend and reuse existing annotations, and automatically generate code for dialogue system modules. Advantages and possible trade-offs are discussed.

Keywords: interoperable annotations, dialogue acts, application semantics

1. Introduction

In context-update approaches to dialogue modelling, a *dialogue act* has two components: a *semantic content*, which describes the objects, properties, relations, or actions that the dialogue act is about, and a *communicative function*, which specifies how an addressee should update their information state with the semantic content. From 1980s, a number of dialogue act annotation schemes has been developed, ranging from simple lists of mutually exclusive tags to complex multi-layered taxonomies. Either used for the analysis of dialogue phenomena or to design dialogue systems, dialogue act annotation has for the most part been limited to marking up communicative functions.

In 2012, the ISO 24617-2 dialogue act annotation standard has been released, which presents a comprehensive multidimensional annotation scheme. The standard was also focused mostly on annotation of communicative functions, however, introduced the notion of type of semantic content - *dimension* - as a shallow characterisation of semantic content of the performed act, i.e. particular type of information state that is updated (ISO, 2012a). The annotation of semantic content is optional, since only task-related acts have full-fledged domain-specific semantic content, while dialogue acts performed for the purpose of dialogue control have marginal semantic content; the meaning of such a dialogue act is concentrated in its communicative function and dimension. In ISO 24617-2 2nd Edition (2019), a protocol is proposed to specify and integrate annotations of semantic content into dialogue act annotations as a 'plug-in',

linking structures of the host annotation scheme to those of the plug-in scheme, see (Bunt, 2019).

Since a single annotation scheme that fully specifies the meaning of natural language dialogue contributions and has sufficient expressive capabilities to build efficient applications is challenging and maybe even not desirable for practical reasons, we deal in practice with multiple existing and newly defined annotation schemes that address different aspects of utterance meaning. Aiming to achieve an adequate coverage of the application-specific semantic content of dialogue acts, in this paper we present an approach that combines a general domain-independent scheme to represent annotations of functional aspects of dialogue contributions (viz. the ISO 24617-2 Dialogue Act Markup Language, DiAML) with multiple possible annotation schemes for representing application-specific semantic content.

Annotation efforts are labour intensive, therefore practical considerations along with theoretical clarity and soundness are important. Multiple schemes can be imported and included, and transformed to be re-usable for certain classes of applications. The schemes need to be explicitly defined and decisions concerning their fusion should be made prior to their use for annotation. The more explicitly dialogue act components are defined, the higher the interoperability level can be achieved and the more robust dialogue applications can be developed. The ISO 24617-2 standard does not prescribe content annotation schemes to be defined directly within a specific annotation design effort. In the following sections we consider the design of annotation schemes for applications of

various semantic complexity, discussing three use cases, and introduce the methodology for their integration with DiAML.

This paper is structured as follows. Section 2 discusses the DiAML annotation scheme and its XSD-based architecture. Section 3 discusses semantic content specifications for scenarios of various complexity, representing (1) intent and slot-filling; (2) term-based information retrieval; and (3) elaborate situation and experience modelling. Section 4 deliberates on the value of the proposed design for real-world applications, discussing advantages and possible trade-offs for system design and annotation work, its costs and its quality. Section 5 concludes the paper with observations on the experiences reported in preceding sections, and outlines directions for future development.

2. DiAML

The ISO 25617-2 dialogue annotation scheme has been designed according to the ISO principles of semantic annotation (Bunt, 2015) and has a three-part definition consisting of (1) an abstract syntax specifying the possible *annotation structures* as set-theoretical constructs; (2) a semantics specifying the meaning of the annotation structures defined by the abstract syntax; (3) a concrete syntax which specifies a representation format for annotation structures.

2.1. Abstract Syntax

The abstract syntax specifies a store of basic concepts, called the ‘conceptual inventory’. The DiAML conceptual inventory consists of:

- a set of dimensions;
- a set of communicative functions;
- a set of qualifiers;
- a set of semantic and pragmatic relations for relating dialogue acts within a dialogue;
- a set of dialogue participants;
- primary data, segmented into markables.

Given a conceptual inventory, the abstract syntax specifies certain pairs, triples, and more complex nested structures made up from the elements of the inventory. Two types of structure are distinguished: *entity structures* and *link structures*. An entity structure contains semantic information about a segment of primary data, and is formally a pair $\langle m, s \rangle$ consisting of a markable and certain semantic information. A link structure contains information about the way segments of primary data are semantically related.

Formally, an entity structure in DiAML is a pair $\langle m, \langle S, A, H, D, F, E, Q \rangle \rangle$ consisting of a markable and a functional dialogue act structure, which is made up by seven components: (1) a sender (S),

(2) one or more addressees (A), (3) zero or more other participants (H), (4) a dimension (D), (5) a communicative function (F), (6) zero or more dependence relations to a set (E) of other dialogue acts, and (7) zero or more qualifiers (Q), where the components H , E , and Q are not necessarily present.

A link structure in DiAML is a triple $\langle e, E, R \rangle$ consisting of an entity structure e , a set of entity structures E , and a relation R .

A full-blown annotation structure for a dialogue in DiAML is a set of entity (e_i) structures and (link (L_j)) structures $\{e_1, \dots, e_n, L_1, \dots, L_k\}$.

2.2. Semantics

The DiAML semantics consists of the specification of a recursive interpretation function I_{DA} which, applied to a semantic content, forms an information state update operation. The DiAML semantics is compositional in the sense that the interpretation of an annotation structure is obtained by combining the interpretations of its component entity and link structures, see (Bunt, 2014) for details.

Semantic issues in using annotations from multiple schemes are addressed in (Bunt, 2024).

2.3. Concrete Syntax

The annotation structures defined by the DiAML abstract syntax can be represented in a variety of semantically equivalent ways, which can encode the structures of the abstract syntax. The official DiAML specification as part of the ISO 24617-2 standard includes a reference representation format based on XML.

For the representation of entity structures an XML element `<dialogueAct>` is defined, with an attribute `@xml:id` whose value is a unique identifier; an attribute `@target`, whose value anchors the annotation in the primary data; and the following attributes: `@sender`, `@addressees`, `@other participants` (optional), `@dimension`, `@communicative function`, `@dependences` (optional), and `@qualifiers` (optional).

The XML elements `<rhetoricalLink>` and is defined for expressing representing rhetorical (‘pragmatic’) relations between dialogue acts, with the attributes `@dact`, `@rhetorelatum`, and `@relType`.

2.4. XSD Definition and Use

DiAML definitions are specified in the form of XSD schema files. XSD schemes provide ‘namespaces’ as a scoping mechanism for XML across multiple schemes and support their integration, inclusion and transformation. Unfortunately, neither homogeneous, nor heterogeneous nor chameleon

design patterns (Costello, 2006; Ko and Yang, 2017) can fulfill the requirements of complex interoperable semantic annotation design. We propose a mixed-patterns approach that steers developers towards a clear data organization and the interoperability of annotations; in addition it enables formal validation of XML documents and automatic code generation to represent and use data from those XML documents inside an application. Code generation is important from a practical point of view; it supports the design of applications that are based on standard interoperable annotations.

The main definitions of the DiAML standard are stated in `DiAML_Types.xsd` and defined within `diaml` namespace. Auxiliary definitions are namespace-less concepts and are defined in `DiAML_Containers.xsd`. The main element `<DialogueAct>` is defined as follows¹:

```
(1) <xs:schema
  targetNamespace=
    "http://www.iso.org/diaml"
  xmlns:xs="http://www.w3.org/2001/
    XMLSchema"
  xmlns:diaml="http://www.iso.org/
    diaml">
  ...
  <xs:complexType name="DialogueAct">
  <xs:attribute ref="xml:id"
    use="required"/>
  ...
</xs:complexType>
</xs:schema>
```

The semantic content of a dialogue act is defined at application level. For this purpose the `<dialogueAct>` element is re-defined in the application-specific scheme, e.g. for the DBOX project²:

```
(2) <xs:schema
  targetNamespace=
    "http://www.dbox.eu/content_spec"
  xmlns:xs=
    "http://www.w3.org/2001/XMLSchema"
  xmlns:diaml="http://www.iso.org/diaml"
  xmlns:dbox=
    "http://www.dbox.eu/content_spec"
  >
  ...
  <xs:element name="dialogueAct">
  <xs:complexType>
  <xs:complexContent>
  <xs:extension
```

¹Note that here and elsewhere in the text XSD and XML examples are excerpts from complete schemes and documents, for reasons of space.

²<https://www.lsv.uni-saarland.de/past-projects/d-box/>

```
base="diaml:DialogueAct">
<xs:sequence>
<xs:element
  ref="dbox:semanticContent"
  minOccurs="0" maxOccurs="1"/>
</xs:sequence>
</xs:extension>
</xs:complexContent>
</xs:complexType>
</xs:element>
...
</xs:schema>
```

For convenience, `DiAML_Containers.xsd` scheme is included without a namespace:

```
(3) <xs:include
  schemaLocation="DiAML_Containers.xsd"/>
```

The scheme contains definitions of elements such as `<diaml>` which in turn contain sequences of dialogue acts and possibly some other elements:

```
(4) <xs:element name="diaml">
  <xs:complexType>
  <xs:sequence>
  <xs:element ref="dialogueAct"
    minOccurs="1"
    maxOccurs="unbounded"/>
  </xs:sequence>
  </xs:complexType>
</xs:element>
```

Note that `DiAML_Containers.xsd` refers to `<dialogueAct>` elements defined in the application-specific scheme and not to the `<DialogueAct>` type in `DiAML_Types.xsd`. Since Containers have no namespace they are placed into the application namespace, e.g. `dbox`, following the *chameleon namespace design*.

The proposed architecture allows to bypass forward-referencing from Containers to application-specific `<dialogueActs>` with application-specific semantic content, while maintaining XML document verification and XML bindings code generation. The semantic content of a dialogue act is defined outside `diaml` and is represented as `<SemanticContent>` elements in the corresponding application-specific XSD, see examples in Section 3. Similar to `DiAML_Types.xsd`, further relevant schemes specifying semantic content can be included, for example, those developed with the ISO Semantic Annotation Framework (SemAF), see (Pustejovsky and Ide, 2017). This follows the *heterogeneous design pattern*.

Dialogue acts are included into `DiAML_Containers.xsd` in: (1) dialogue annotations defining participants, tokens, sounds

and functional segments; (2) a corpus which consists of (3) dialogue sessions with reference to segmented primary data; and (4) messages exchanged between dialogue system modules.

Containers are supporting types and are not obligatory to use. However, they help maintain annotation consistency, and serve as examples to design one's own Containers, for instance, when different primary data representation formats are desired or other types of annotations need to be performed.

The architecture enables formal validation with standard tools like Oxygen³ and automatic code generation with, for example, Java XML bindings (JAXB). Consider examples in (5) for automatically generated Java code for dialogue act and in (6) for a <diaml> element:

```
(5) @XmlElement(name =
    "dialogueAct")
    public class DialogueAct
    extends org.iso.diaml.DialogueAct
    {
    protected DboxSemanticContent
        DboxSemanticContent;
        /* Gets the value of the
        * DboxSemanticContent property.
        * @return possible object is
        * @link DboxSemanticContent
        */
    public DboxSemanticContent
        getDboxSemanticContent() {
    return DboxSemanticContent;
    }

        /* Sets the value of the
        * DboxSemanticContent property.
        * @param value allowed object is
        * {@link DboxSemanticContent }
        */
    public void setDboxSemanticContent(
        DboxSemanticContent value)
    {
        this.DboxSemanticContent = value;
    }
    }
```

```
(6) @XmlElement(name = "diaml")
    public class Diaml {
    @XmlElement(required = true)
        protected List<DialogueAct>
        dialogueAct;
    /**
        * Gets the value of the
        * dialogueAct property.
        * This accessor method
        * a reference to the live
```

```
* list, not a snapshot. Therefore
* any modification you make to the
* returned list will be present
* inside the JAXB object.
*/
public List<DialogueAct>
    getDialogueAct() {
    if (dialogueAct == null) {
    dialogueAct = new
    ArrayList<DialogueAct>();
    }
    return this.dialogueAct;
    }
}
```

It may be observed that the automatically generated code in (5) and (6) strictly follows XML element definition patterns from the specified XSD schemes. More specifically, dbox DialogueAct class extends diaml DialogueAct class by adding the dbox semanticContent field. dbox Diaml class contains a field of type List<DialogueAct>, where <DialogueAct> refers to dbox <DialogueAct> and not diaml <DialogueAct>, i.e. Diaml is the list of application-specific (dbox) dialogue act types, rather than generic diaml dialogue acts.

3. Application Schemes: use cases

This Section presents XML Schemes which capture semantics of the problem domain (Application Semantics) at conceptual level and represent it in XML schema definition language (XSD). We proceed from simple to more complex schemes featuring real use case scenarios.

3.1. Intents and Slot Filling

In the past few years, conversational AI agents have become extremely popular. Traditional conversational agents are often modeled based on *intents*,⁴ which refers to the primary goal of a dialogue utterance. Intents are typically identified by analyzing the words and phrases in an utterance and mapping them to predefined categories or concepts. For example, an utterance like "What time are there trains from Norwich to York?" might be mapped to an intent like `request_departureTime` where the first part corresponds to the communicative function of an utterance and the second part specifies a high-level semantic content, e.g. 'topic'. Additional entities are extracted to refine, modify and provide more context to the intent. For example, 'Norwich' and 'York', and specified

³https://www.oxygenxml.com/xml_editor.html

⁴Currently, intentless agents are claiming the ground, see <https://rasa.com/blog/breaking-free-from-intents-a-new-dialogue-model>.

as slot types of `departure_location` and `destination_location` respectively. Many task-oriented information-seeking dialogues are modelled this way (Larson and Leach, 2022).

Contrary to the traditional two-component intent definition, we break up intent specifications into two schemes: (1) a DiAML representation for a functional component; and (2) an Application Scheme for a semantic content for a particular domain. For example, the DBOX dialogues collected to design Question Answering Dialogue System (QADS) are modelled using this approach. Players ask questions about biographical facts of an unknown person in order to guess their identity. Questions are classified with their communicative function (e.g. Propositional, Check, Set and Choice Questions) and semantic content based on the Expected Answer Type (EAT). For the latter, 59 semantic relations between entities (e.g. between participants or between an event and participants) have been defined extending the Knowledge Base Population Slot Filling Task (TAC KPB, Min and Grishman (2012)). Each relation has two arguments and is one of the following types:

- $\text{RELATION}(Z, ?X)$, where Z is the person in question and X the entity slot to be filled, e.g. $\text{CHILD OF}(\text{einstein}, ?X)$;
- $\text{RELATION}(E1, ?E2)$ where $E1$ is the event in question and $E2$ is the event slot to be filled, e.g. $\text{REASON}(\text{death}, ?E2)$; and
- $\text{RELATION}(E, ?X)$ where E is the event in question and X the entity slot to be filled, e.g. $\text{DURATION}(\text{study}, ?X)$.

The slots are categorized by the content and quantity of their fillers. Slots are labelled as name (person, organization, or geo-political entity), value (a numerical value or a date), or string. Slots can be as single-value (e.g. date of birth) or list-value (e.g. employers) based on the number of fillers they can take (Petukhova et al., 2018). Consider an excerpt from the DBOX XSD scheme:

```
<xs:schema
xmlns:xsd=
  "http://www.w3.org/2001/XMLSchema"
targetNamespace=
  "http://www.dbox.eu"
xmlns="http://www.dbox.eu"
elementFormDefault="unqualified">
<xs:import namespace=
  "http://www.iso.org/diaml"
schemaLocation="DiAML_Types.xsd"/>
<xs:include schemaLocation=
  "DiAML_Containers.xsd"/>
<xs:simpleType name="eatRelation">
  <xs:restriction base="xs:token">
  <xs:enumeration value="origin"/>
  ...
  <xs:enumeration value="locBirth"/>
```

```
</xs:restriction>
</xs:simpleType>
...
<xs:simpleType name="SlotFiller">
  <xs:restriction base="xs:token">
  <xs:enumeration value="name"/>
  ...
  </xs:restriction>
</xs:simpleType>
...
<xs:simpleType name="GPE">
  ...
</xs:simpleType>
</xsd:schema>
```

A simple representation of semantic content can be defined as a list of attribute-value pairs as in 7.

(7) Player (P1): What country are you from?
System (P2): US

```
<dialogueAct xml:id="dap1TSK0"
  sender="#p1" addressee="#p2"
  dimension="task"
  communicativeFunction="setQuestion"
  target="#fsp1TSKCV0">
  <dbox:semanticContent>
    <entity xml:id="x1" target="#ne1"
      type="name" value="person"
      quantity="single"/>
    <entity xml:id="x2" target="#ne2"
      type="name" value="GPE"
      quantity="single"/>
    <eatRelation source="#x1"
      slotFiller="#x2" type="origin"/>
  </dbox:semanticContent>
</dialogueAct>

<dialogueAct xml:id="dap2TSK1"
  sender="#p2" addressee="#p1"
  dimension="task"
  communicativeFunction="answer"
  target="#fsp2TSKCV1"
  functionalDependence="#dap1TSK0">
  <dbox:semanticContent>
    <entity xml:id="x1" target="#ne1"
      type="name" value="person"
      quantity="single"/>
    <entity xml:id="x2" target="#ne2"
      type="name" value="US"
      quantity="single"/>
    <eatRelation source="#x1"
      slotFiller="#x2" type="origin"/>
  </dbox:semanticContent>
</dialogueAct>
```

Player asks the question concerning the *country* (markable $x2$, named entity $ne2$) of origin of the person in question (markable $x1$ assigned to *you*, named entity $ne1$). We expect an answer of relation type $\text{ORIGIN}(x1, ?x2)$ where $x1$ is the person whose identity need to be guessed and $x1$ the entity slot to be filled. A single slot filler is expected of type GPE, filled in answer with 'US'.

3.2. Term-based Information Retrieval

The specification of semantic content may include elements from external knowledge bases or ontologies. For example, as a use case, we simulated pre-operative question answering sessions between doctors and patients. As a core part of these medical encounters, Patient Education Forms (PEFs) have to be filled in, and the patient's informed consent form signed. It is of chief importance that the forms are properly understood, and that medical procedures and risks are explained. PEFs contain many medical terms including some in Latin and some as abbreviations. These terms have to be detected and corresponding definitions retrieved from available medical documents. Thus, our approach was to detect medical terms, map them to entries of existing databases and ontologies, and retrieve definitions. For more information concerning the term extraction and application details see (Wolf et al., 2019; Bhatt, 2022).

There is a range of medical knowledge bases, ontologies, standard terminologies, and lexicons. One of the most widely used repositories of biomedical terms is the Unified Medical Language System (UMLS⁵, Bodenreider (2004)), which integrates over 2 million names for 900 000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts. We used MetaMap⁶ to find UMLS Metathesaurus concepts and to generate lexical variants of concept names. MetaMap gives a relevance score to each concept. In UMLS, similar terms (`biomedicalTerm`) from different vocabularies are grouped into the same concept (`umlsConcept`) and receive a Concept Unique Identifier (`umlsCUI`). Terms are grouped into semantic groups (`umlsSG`) and semantic types (`umlsST`) through which synonyms and related terms can be accessed. One of the vocabularies integrated into UMLS, which is frequently used for text simplification, is the Consumer Health Vocabulary (CHV, Zeng et al. (2007)), which comprises terms (`chvTerm`) for many common words and phrases used by health care professionals. Another frequently used vocabulary is SNOMED CT, Benson (2012) which consists of a large number of concepts (`snomedctConcept`) from clinical reports. We identified terms related to PEFs in SNOMED CT (84.9% - 95.9%) and in CHV (73.5% - 80.0%). Definitions were mostly retrieved from MedlinePlus⁷, an online public health information resource (Schnall and Fowler, 2013).

UMLS concepts, matching CHV and SNOMED CT terms, are integrated with retrieved Med-

linePlus definitions as part of semantic content of the BRENNDA (Business pROcess modElS iNtegration iNto Dialogue mANagement) system (Tarakameh, 2019):

```
<xs:schema
xmlns:xsd=
  "http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.brennda.org"
xmlns="http://www.brennda.org"
elementFormDefault="unqualified">
<xs:import
  namespace="http://www.iso.org/diaml"
  schemaLocation="DiAML_Types.xsd"/>
<xs:include
  schemaLocation="DiAML_Containers.xsd"/>

<xsd:complexType
name="semanticContent">
  <xsd:sequence>
    <xsd:element name="umlsConcept"
      <xs:attribute umlsCUI="xml:id"
        use="required"/>
    <xsd:element name="snomedctConcept"
      <xs:attribute SCTID="xml:id"
        use="optional"/>
    <xsd:element name="chvTerm"
      <xs:attribute SCUI="xml:id"
        use="optional"/>
    <xsd:element
      name="medlineplusDefinition"
      <xs:attribute
        health-topicID="xml:id"
        use="required"/>
  </xsd:sequence> </xsd:complexType>
</xsd:schema>
```

The application of this approach gives rise to the following dialogue fragment:

- (8) Patient (P1): What is sleep apnea?
System (P2): It is a sleep-disordered breathing

```
<dialogueAct xml:id="dap1TSK13"
sender="#p1" addressee="#p2"
dimension="task"
communicativeFunction="setQuestion"
target="#fsp1TSKCV13"
  <brennda:semanticContent>
    <biomedicalTerm xml:id="biot21"
target="#ne21"
umlsConcept="sleepApnea"/>
  </brennda:semanticContent>
</dialogueAct>

<dialogueAct xml:id="dap2TSK27"
sender="#p2" addressee="#p1"
dimension="task"
communicativeFunction="answer"
target="#fsp2TSKCV27"
functionalDependence="#dap1TSK13"
  <brennda:semanticContent>
    <biomedicalTerm xml:id="biot21"
```

⁵<https://www.nlm.nih.gov/research/umls/>

⁶<https://www.metamap.com>.

⁷<https://medlineplus.gov/>


```

target="#ne21"
umlsConcept="sleepApnea"
umlsCUI="C0018787"
umlsSG="disorder"
umlsST="diseaseOrSyndrom"
chvTerm="sleep
  -disordered breathing"
snomedctConcept="sleepApnea
  disorder"
medlineplusDefinition="sleep
  -disordered breathing"/>
</brennda:semanticContent>
</dialogueAct>

```

In (8), patient (P1) filling in the PEF has difficulty to understand what ‘sleep apnea’ is. The system extracts the term from the patient’s question (*ne21*) and queries the UMLS ontology for CHV and SNOMED CT concepts (synonyms) and the MedlinePlus definition. If a term was found in CHV, it was considered unnecessary to provide other synonyms in a generated answer.

3.3. From Situations to Experiences

Other applications may require richer semantic content to be incorporated in a dialogue semantics than illustrated above. This is, for example, the case of situated (or context-aware) human-computer interactions involving multiple human and artificial participants with certain properties performing various roles, dealing with, referring to, and reasoning about the world within a certain environment/context engaged in various events that take place in a certain time and space. The more complex the situation, the richer the content specification is required to describe it.

There have been numerous attempts to define context-aware interactions, most of which are very specific and provide too limited support for situation abstraction. Fully specified semantic representation is hardly possible, and sometimes not even desirable and feasible for maintaining experimental control. One of the most recently undertaken attempts to design annotations which support whole-sentence semantic representation is the Abstract Meaning Representation initiative (AMR, Banarescu et al. (2013)), with an extension for dialogue semantics (Dialogue-AMR, Bonial et al. (2020)). Within this framework, human-robot dialogues are annotated with a speech act⁸, tense (*before, now, after*), aspect (*stable, ongoing, complete, habitual, completable*) and semantic role information (PropBank, Palmer et al. (2005)); see example (9) adopted from (Bonial et al., 2020).

⁸The designed tagset to model human-robot interactions comprises a list of 14 mutually exclusive tags.

```

(9) Commander (C2): Drive to the door
(c / command-SA
  :ARG0 (c2 / commander)
  :ARG2 (r / robot)
  :ARG1 (g / go-02 :completable +
  :ARG0 r
  :ARG3 (h / here)
  :ARG4 (d/ door)
  :time (a2 / after
  :op1 (n / now)))

```

To parse and generate DialAMRs, AMR parsers and resources are used, which are steadily growing in number and scope (Zhou et al., 2021; Cheng et al., 2022; Vasylenko et al., 2023).

In the past few years, a number of ISO SemAF annotation schemes have been developed, besides DiAML: Time and Events (ISO, 2012b), Semantic Roles (ISO, 2014), Semantic Relations in Discourse (ISO, 2016), Coreference (ISO, 2019b), Spatial Information (ISO, 2019a) and Quantification (ISO, 2019c). It would be very attractive to include these schemes for modeling situated interactions. An elegant way to incorporate SemAF annotations into dialogue act annotations has been proposed by Bunt (2019), using *annotation schema plug-ins* which make use of a variety of content link structures, e.g. *contentLink* and *emoLink*, for importing elements of one annotation schema into another. Multiple SemAF schemes can be used for content representation by means of the *interlinking* technique (Bunt, 2024).

Dialogue participation involves a range of social and emotional experiences. Human interactions are more than the exchange of information, decision making, or problem-solving; they involve a wide variety of aspects related to feelings, emotions, social status and interpersonal relations.

For developing socially embedded dialogue systems, it has been proposed to model interactive behaviour in terms of *experiences*, i.e. instances of mental states or dialogue context/states (Stevens et al., 2016; Malchanau et al., 2018). Dialogue participants collect interactive experiences and learn from them. An instance may encode all information that influences the interpretation and generation of dialogue contributions, and thus the decision making process: knowledge about domain and partners, participants’ preferences and attitudes, emotional state and social status, and this list is far from exhaustive. Although there are no theoretical limitations on instance size, the application efficiency is the highest when the state representation is relatively compact. A very complex state representation may make state tracking and instance retrieval very costly. There should be no problem with using incomplete instances, since humans also have to deal with partially available, am-

| Holder: slot type | Possible Values |
|-------------------------|---|
| doctor: strategy | competence warmth |
| doctor: expertise | low moderate high |
| doctor: importance | low moderate high |
| doctor: framing-effect | threat risk benefit |
| doctor: preference | (im)possible (un)desired (in)abile mandatory urgent |
| patient: strategy | avoiding hesitant submissive biased cooperative aggressive resistant |
| patient: expertise | low moderate high |
| patient: importance | low moderate high |
| patient: framing-effect | threat risk benefit |
| patient: preference | (im)possible (un)desired (in)abile mandatory urgent |
| patient: readiness | low moderate high |

Table 1: Instance contents concerning participants’ strategies and preferences.

biguous and/or vague information, imperfect understanding and limitations of working memory.

We designed instance-based LICA⁹ agents that are involved in doctor-patient interactions, where an imbalance is observed in the knowledge and relationship between interlocutors, due to social, professional and personal factors. Agents simulate patients of different personalities, motivational and emotional dispositions. Interacting with LICA agents, doctors are trained to identify strategies that are optimal for specific patients, i.e. positively affect patient’s preferences for a certain treatment.

Some important strategies concern pragmatic aspects such as use of indirect speech acts for politeness or to express interest, respect, support and empathy; or qualified functional aspects concerning affected behaviour in order to build a trustful relationship through the development of rapport and responsiveness to a patient’s emotions (‘Appeal to Warmth’, (Fiske, 2018). Other strategies concern the quality of arguments presented in health intervention utterances (‘Appeal to Competence’): (1) information provided, e.g. expert language use and appeal to authority; (2) attitudes towards proposed interventions and its outcomes: costs, appeal to importance and call for readiness; and (3) targeted framing effects, e.g. presentation of options in positive or negative terms (survival rates or mortality rates for a treatment).

Both relevant functional aspects and semantic content are encoded in an instance represented as a set of slot-value pairs. Table 1 presents a template encoding beliefs concerning domain knowledge, the participants’ preferences, and the persuasion strategy being pursued.

The domain selected for our use case concerns the treatment of diabetes. To generate health interventions of various types, medical claims and evidence were collected from PubMed abstracts¹⁰, viz. 32 claims and 64 supporting and attacking evidence statements. Keywords

⁹Learning Intelligent Conversational Agents.

¹⁰<https://pubmed.ncbi.nlm.nih.gov/>

and phrases were extracted using the KeyBERT model, (Grootendorst, 2020) and the term banks UMLS and CHV were queried to compute the level of expertise, the framing effects, and the applied strategy. Importance, readiness, preference and framing effects were modulated. On the basis of previous research (Guenoun and Zlatev, 2023; Lapina and Petukhova, 2017), features were selected for linguistic modulations. These concern *appeal* (competence/warmth), *text length* (long/short), *framing* (risk/benefit), *lexical complexity* (complex/simple), *concreteness* (numbers/textual delivery) and *grammatical voice* (passive/active) (Wan Ching Ho and Petukhova, 2024).

Below is an excerpt from the XSD scheme specifying LICA semantic content; a dialogue fragment example which makes use of LICA content specifications is presented in (10) of Appendix 7:

```
<xs:schema
xmlns:xsd=
  "http://www.w3.org/2001/XMLSchema"
targetNamespace="http://www.lica.org"
xmlns="http://www.lica.org"
elementFormDefault="unqualified">

  <xs:import
    namespace="http://www.iso.org/diaml"
    schemaLocation="DiAML_Types.xsd"/>
<xs:include schemaLocation=
  "DiAML_Containers.xsd"/>

  <xs:simpleType name="Holder">
    <xs:restriction base="xs:token">
      <xs:enumeration value="doctor"/>
      <xs:enumeration value="patient"/>
    </xs:restriction>
  </xs:simpleType>

  <xs:simpleType name="Strategy">
    <xs:restriction base="xs:token">
      <xs:enumeration value="competence"/>
      ...
      <xs:enumeration value="hesitant"/>
    </xs:restriction>
  </xs:simpleType>
  ...
  <xs:simpleType name="Readiness">
    <xs:restriction base="xs:token">
      <xs:enumeration value="low"/>
      <xs:enumeration value="moderate"/>
      <xs:enumeration value="high"/>
    </xs:restriction>
  </xs:simpleType> </xsd:schema>
```

4. Value for Real-World Applications

The multi-scheme design presented in this paper has a number of advantages, as well as limitations. One of the advantages is that such a design splits up large annotation efforts into small(-

er) tasks that are more manageable for human annotators and automatic labeling systems. This positively affects annotation quality and costs: it increases annotation consistency and accuracy, it improves scheme usability in terms of inter-annotator agreement, and it potentially decreases annotation time¹¹.

Another advantage is that task-specific annotations can be straightforwardly reused by other applications. For instance, labeled data can be used for adaptation or knowledge distillation of pre-trained large models, which significantly improves their performance on a variety of up-/downstream tasks. Applications based on clear use-case semantics are easier to evaluate and their performance can be directly compared to other existing systems or models.

A limitation of this approach is that semantic information that is not captured in annotations needs to be modelled inside an application and often remains somewhat hidden. This is for instance the case of hidden layers as used in modern neural systems that are responsible for learning intricate structures in data which are not explicitly annotated. This makes neural networks a powerful but black-boxed tool with limited explainability and interpretability of the system's behaviour. Another limitation is that the collection of semantic information from multiple annotation projects runs into danger to be less interoperable and challenging to fuse.

Advantages and limitations put the designer in a position to carefully weight pros and cons for their design scenarios, with trade offs between semantic expressiveness and precision on the one hand and simplicity of its application on the other.

5. Discussion and Conclusion

This paper explores dialogue use cases of varied semantic complexity: slot-filling supporting question answering, term-based information retrieval, and complex situation and experience specifications. Functional aspects of dialogue contributions are modelled using the ISO 24617-2 dialogue act annotation standard and specified in DiAML. Semantic content is represented in an appropriate way for a specific dialogue application.

Applications require an interpretation framework, either utilising explicit knowledge representation techniques or relying on an intuitive interpretation scattered implicitly across application code. Specifying annotation schemes for semantic content in a formal way, e.g. in XSD format, opens

¹¹For example, the Real Time Factor (RTF) can be estimated - amount of time spent on annotations given the amount of dialogue data. RTF 10 means that an annotator spent 100 minutes annotating 10 minutes of real dialogue, e.g. speech and video.

opportunities to share annotations among different applications and tools.

In this paper we have proposed a way to integrate a wide range of domain/application-specific annotations with the domain-independent ISO 24617-2 scheme specified in DiAML. The ISO annotation standards developed within SemAF can be integrated in a similar manner. For all components, XML schema definitions (XSD) refer to external XML schemes. More than one XML schema can be included or imported within an XML schema, as we showed using 'namespaces'. XSD has the important advantage that it can be used to validate the contents of an XML document, as well as to generate code within an application design.

We will distribute the designed XSD domain-independent DiAML and Application Schemes on the DialogBank¹², a collection of dialogues annotated according to ISO 24617-2 standard. A full package of gold standard dialogue act annotations, XSD schemes, primary data, and documentation is available for the Metalogue Multi-Issue Bargaining Corpus in the LDC catalogue.¹³

6. Acknowledgments

The authors are also very thankful to anonymous reviewers for their valuable comments.

7. Bibliographical References

- Laura Banarescu et al. 2013. Abstract meaning representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*.
- Tim Benson. 2012. *Principles of health interoperability HL7 and SNOMED*. Springer Science & Business Media.
- Abhinav Bhatt. 2022. Personalized medical arguments generation. Master's thesis, Saarland University.
- Olivier Bodenreider. 2004. The Unified Medical Language System (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.
- Claire Bonial et al. 2020. Dialogue-AMR: abstract meaning representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

¹²<https://dialogbank.lsv.uni-saarland.de/>

¹³<https://catalog.ldc.upenn.edu/LDC2017S11>

- Harry Bunt. 2014. Annotations that effectively contribute to semantic interpretation. In *Computing Meaning: Volume 4*, pages 49–69. Springer.
- Harry Bunt. 2015. On the principles of semantic annotation. In *Proceedings of the 11th Workshop on Interoperable Semantic Annotation (ISA-11)*.
- Harry Bunt. 2019. Plug-ins for content annotation of dialogue acts. In *15th Workshop on Interoperable Semantic Annotation (ISA-15)*.
- Harry Bunt. 2024. Combining semantic annotation schemes through interlinking. In *20th Workshop on Interoperable Semantic Annotation (ISA-20)*.
- Ziming Cheng, Zuchao Li, and Hai Zhao. 2022. Bibl: AMR parsing and generation with bidirectional bayesian learning. In *29th International Conference on Computational Linguistics*.
- R.L. Costello. 2006. [MI schemas: Best practices. multi-schema project: Zero, one, or many namespaces?](#)
- Susan T Fiske. 2018. Stereotype content: Warmth and competence endure. *Current directions in psychological science*, 27(2):67–73.
- Maarten Grootendorst. 2020. [Keybert: Minimal keyword extraction with bert.](#)
- Bushra S Guenoun and Julian J Zlatev. 2023. Sending signals: Strategic displays of warmth and competence. *Working Paper 23-051*.
- ISO. 2012a. *Language resource management – Semantic annotation framework – Part 2: Dialogue acts. ISO 24617-2*. ISO Central Secretariat, Geneva.
- ISO. 2012b. *Language Resource Management-Semantic Annotation Framework (SemAF)- Part 1: Time and events. ISO 24617-1*. ISO Central Secretariat, Geneva.
- ISO. 2014. *Language Resource Management- Semantic Annotation Framework (SemAF)- Part 4: Semantic roles ISO 24617-4*. ISO Central Secretariat, Geneva.
- ISO. 2016. *Language resource management - Semantic annotation framework (SemAF) Part 8, Semantic relations in discourse ISO 24617-8*. ISO Central Secretariat, Geneva.
- ISO. 2019a. *Language resource management - Semantic annotation framework- Part 7: Spatial information (ISO-Space). ISO 24617-7*. ISO Central Secretariat, Geneva.
- ISO. 2019b. *Language Resource Management-Semantic Annotation Framework (SemAF)- Part 9: Reference annotation framework (RAF) ISO 24617-9*. ISO Central Secretariat, Geneva.
- ISO. 2019c. *Language Resource Management: Semantic Annotation Framework (SemAF)- Part 12: Quantification. ISO 24617-12*. ISO Central Secretariat, Geneva.
- Hye-Kyeong Ko and Minh Yang. 2017. An effective xml schema conversion technique for improving xml document reusability using pattern list. *International Journal of Internet, Broadcasting and Communication*, 9(2):11–19.
- Valeria Lapina and Volha Petukhova. 2017. Classification of modal meaning in negotiation dialogues. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.
- Stefan Larson and Kevin Leach. 2022. [A survey of intent classification and slot-filling datasets for task-oriented dialog.](#)
- Andrei Malchanau, Volha Petukhova, and Harry Bunt. 2018. Towards integration of cognitive models in dialogue management: designing the virtual negotiation coach application. *Dialogue & Discourse*, 9(2):35–79.
- Bonan Min and Ralph Grishman. 2012. Challenges in the knowledge base population slot filling task. In *LREC*, pages 1137–1142.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.
- Volha Petukhova et al. 2018. Understanding questions and extracting answers: Interactive quiz game application design. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 246–261. Springer.
- James Pustejovsky and Nancy Ide, editors. 2017. *Handbook of Linguistic Annotation*. Springer.
- Janet G Schnall and Susan Fowler. 2013. Medlineplus. gov: quality health information for your patients. *AJN The American Journal of Nursing*, 113(9):64–65.
- Christopher Stevens, Harmen de Weerd, Fokie Cnossen, and Niels Taatgen. 2016. A metacognitive agent for training negotiation skills. In *Proceedings of the 14th International Conference on Cognitive Modeling (ICCM 2016)*.
- Bahar Tarakameh. 2019. Integration of business process models into dialogue management: patient self-assessment use case. Master’s thesis, Saarland University.

Pavlo Vasylenko et al. 2023. Incorporating graph information in transformer-based amr parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, page 1995–2011.

Clara Wan Ching Ho and Volha Petukhova. 2024. Towards generation of personalised health intervention messages. In *Proceedings of the 1st Workshop on Patient-oriented Language Processing (CL4HEALTH), LREC-COLING 2024*.

Martin Wolf et al. 2019. Term-based extraction of medical information: Pre-operative patient education use case. In *Conference on Recent Advances in Natural Language Processing*.

Qing Zeng et al. 2007. Term identification methods for consumer health vocabulary development. *Journal of medical Internet research*, 9 (1).

Jiawei Zhou et al. 2021. Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based AMR parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Appendix: LICA dialogue fragment

- (10) Doctor (P1): You should minimise alcohol intake
Doctor (P1.1): Alcohol intake may place people with diabetes at increased risk for delayed hypoglycemia
Doctor (P1.2): Persons using insulin or insulin secretagogues can experience delayed nocturnal or fasting hypoglycemia after alcohol consumption.
Doctor (P1.3): Moderate alcohol consumption has minimal acute and/or long-term detrimental effects on glycemia with type 1 or type 2 diabetes.

```
<dialogueAct xml:id="dap1TSK1"
sender="#p1" addressee="#p2"
dimension="task"
communicativeFunction="inform"
target="#fsp1TSKCV1"
  <lica:semanticContent>
    <claim xml:id="claim1"
target="#fsp1TSKCV1"
topic="alcohol intake"
preference="mandatory"
    </lica:semanticContent>
</dialogueAct>
```

```
<dialogueAct xml:id="dap1TSK1.1"
sender="#p1" addressee="#p2"
dimension="task"
communicativeFunction="inform"
target="#fsp1TSKCV1.1"
  <lica:semanticContent>
```

```
<evidence xml:id="evidence1.1"
target="#fsp1TSKCV1.1"
relation="#claim1"
stance="support"
topic="alcohol intake"
expertise="moderate"
importance="high"
preference="mandatory"
framing="risk"
strategy="competence"
</lica:semanticContent>
```

```
</dialogueAct>
<rhetoricalLink dact="#dap1TSK1.1"
rhetoRelatum="#dap1TSK1"
relType="justification"/>
```

```
<dialogueAct xml:id="dap1TSK1.2"
sender="#p1" addressee="#p2"
dimension="task"
communicativeFunction="inform"
target="#fsp1TSKCV1.2"
  <lica:semanticContent>
    <evidence xml:id="evidence1.2"
target="#fsp1TSKCV1.2"
relation="#claim1"
stance="support"
topic="alcohol intake"
expertise="high"
importance="high"
preference="mandatory"
framing="risk"
expertise="high"
strategy="competence"
    </lica:semanticContent>
```

```
</dialogueAct>
<rhetoricalLink dact="#dap1TSK1.2"
rhetoRelatum="#dap1TSK1"
relType="justification"/>
```

```
<dialogueAct xml:id="dap1TSK1.3"
sender="#p1" addressee="#p2"
dimension="task"
communicativeFunction="inform"
target="#fsp1TSKCV1.3"
  <lica:semanticContent>
    <evidence xml:id="evidence1.3"
target="#fsp1TSKCV1.3"
relation="#claim1"
stance="support"
topic="alcohol intake"
expertise="high"
preference="mandatory"
importance="moderate"
framing="risk"
expertise="high"
strategy="competence"
    </lica:semanticContent>
```

```
</dialogueAct>
<rhetoricalLink dact="#dap1TSK1.3"
rhetoRelatum="#dap1TSK1"
relType="justification"/>
```

Annotation-Based Semantics for Dialogues in the Vox World

Kiyong Lee

Korea University, Seoul
ikiyong@gmail.com

Abstract

This paper aims at enriching Annotation-Based Semantics (ABS) with the notion of small visual worlds, called the *Vox worlds*, to interpret dialogues in natural language. It attempts to implement classical set-theoretic models with these *Vox worlds* that serve as interpretation models. These worlds describe dialogue situations while providing background for the visualization of those situations in which these described dialogues take place interactively among dialogue participants, often triggering actions and emotions. The enriched ABS is linked to VoxML, a modeling language for visual object conceptual structures (vocs or vox) that constitute the conceptual basis of visual worlds. Each *Vox world* is characterized by a set of visualized situation types, possibly depicted by static pictures or dynamic videos, to interpret dialogues. This paper focuses on annotating and interpreting a few illustrative dialogues for such a small visual world.

Keywords: annotation-based semantics (ABS), partial information, situation types, small visual world, visual object concept structure (vocs, or vox), *Vox world*,

1. Introduction

1.1. Aim and Overview

This paper aims to enrich Annotation-Based Semantics (ABS), proposed by Lee (2020, 2023), with the notion of *small visual worlds* to annotate and interpret dialogues in natural language. Small visual worlds form the *Vox world*, consisting of visual conceptual object structures (vocs or vox) in the modeling language VoxML (Pustejovsky and Krishnaswamy, 2014). These small visual worlds may be forming *scenes* or "*visually perceived situations*" (Barwise, 1989) with formal constructions involving human perceptions of the surroundings in interactive communications or dialogues.

ABS makes two but related uses of a set of small visual worlds. One use is to describe a dialogue situation in which the dialogue participants interact with each other linguistically through verbal exchanges. The other use is to form a background situation à la Barwise and Perry (1983) or bring in the linguistic or world knowledge for interpreting communicative exchanges and the things involved in them. In annotating dialogues for their act types and content, ABS refers to these two *situation types*, one for describing situations and another for providing background for interpreting them.

For example, part of a dialogue transcript "Husband to Wife: Take this." describes a situation in which the husband says to his wife: "Take this." Suppose the wife responded with a smile to her husband by saying "Thanks. Delicious." Then, this response provides a contextual background for inferring that the deictic expression "this" must refer to something edible or potable for tasting while showing the wife's satisfaction with gratitude. Furthermore, a picture or scene showing how such a

dialogue was enacted provides a background situation for interpreting more vividly what is meant by the husband's utterance "Take this." Such a picture depicts a small visual world or part of it.

1.2. Scope, Focus, and Motivation

The scope of the paper is very much restricted in its form for presentation and data for analysis. This is not a formal paper that formulates the key notions rigidly in logico-mathematical terms. It illustrates how a few short dialogues are annotated and interpreted for such a visually perceptible world, the *Vox world* or part of it. The data for analysis is also very restricted to the extent that no statistical justification is presented for the claims made in the paper.

The paper focuses on the complementary roles of dialogue scripts and related images or pictures that I claim depict a small world providing background for the interpretation. It treats very simple dialogues, having only a few words in the utterances, for illustrations while avoiding the treatment of various dialogue act types and dimensions (e.g., task-oriented vs. expressive (of emotions)) (Bunt, 2022).

Dialogues are chosen as specific data for analysis in this paper because they present the most challenging task for natural language processing in at least three respects. First, annotation may work while syntax fails to process because dialogues have many deictic expressions (e.g., "this" as in "Husband to Wife: Try this.") or ellipses (e.g., "Wife to Husband: Thanks. Delicious.") with syntactic variations and aberrations from regular grammar, unlike written text. Second, dialogue acts often trigger the actions of dialogue participants as agents or objects with some other semantic roles and emo-

tions (e.g.: emotive and evaluative as in "Wife to Husband: Thanks. Delicious.") lie involved in the content of the dialogue conveyed. Annotation can easily mark up such actions enacted and emotions expressed by dialogue participants. Third, the interpretation of dialogue contents requires background information, especially in the applicational context of Human-Computer Interactions (HCI) or Human-Object Interactions (HOI) (e.g.: "Husband to Wife: Try this." requires a variety of actions as responses, depending on its context of use). For these reasons, the treatment of dialogue acts and content is well-motivated, challenging, and most interesting as a linguistic task, especially for computational applications. Computers or robots may participate in a dialogue as artificial agents in a computational application.

1.3. Claim, Proposal, and Basic Assumptions

This paper claims that the set-theoretic model structures for interpreting natural language or its logical forms, as in Montague Semantics (Montague, 1974b; Dowty et al., 1981) should be re-envisioned and re-designed. This must be implemented with visualized small worlds or situation types delimited by the visual object conceptual structures that are well-defined, for instance, by the modeling language VoxML.

In VoxML, in contrast, each object, action, or relation is a first-class citizen in a small world, as proposed in Situation Semantics (Barwise, 1989), that forms a visual object conceptual structure. These structures are then represented by a complex attribute-value matrix (AVM) structure with embedded AVM's that carry a variety of relevant information. Likewise, various types of relations in an interpretation model are defined similarly.

Figure 1 shows how the Annotation-Based Semantics (ABS) is linked to VoxML, [i] linguistically supporting it. ABS annotates communicative language segments including dialogues, [ii] generating annotation structures \mathbf{a} while referring to the voxicon V of VoxML. It then translates annotation structures to logical forms $\sigma(\mathbf{a})$ in typed first-order logic [iii]. These logical forms are then interpreted with respect to the minimal models constrained by the habitats, affordances, and embodiments of denotative elements. For such processes, VoxML as a modeling language introduces Voxicon to list the voxemes that augment the Generative Lexicon (Pustejovsky and Batiukova, 2019) with the notions of Habitat theory (Pustejovsky, 2013) and Gibsonian affordance structures (Gibson, 1977, 1979). These voxemes are represented in complex feature structures, in which some features (attributes) have feature structures as values, as illustrated in

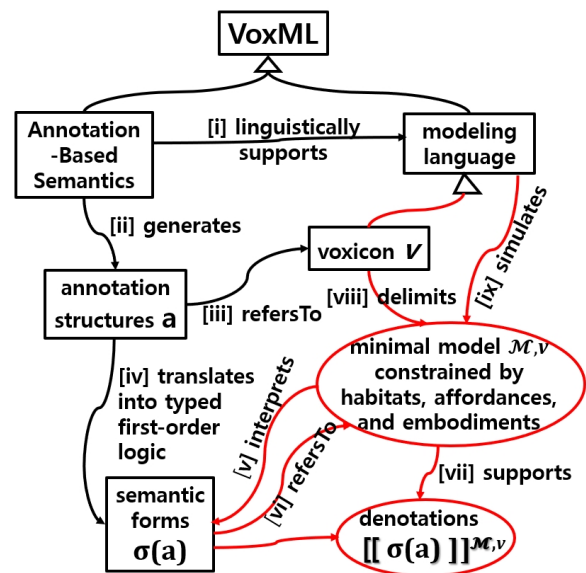


Figure 1: VoxML-linked ABS

Section 5.

This paper claims that the ABS thus designed is linked to VoxML, which constitutes the structural basis of visual worlds. Its sub-module, the *Voxicon*, which comprises *voxemes*, characterizes various visual object concept structures in specific forms. Each Vox world, composed of these structures, is represented by a set of visualized situation types, possibly accompanied by static images or dynamic videos (motion pictures) to interpret dialogues.

As in the Situation Semantics of Barwise and Perry (1983) and Barwise (1989), the *partiality* of information is a foundational notion for ABS. Annotation targets the particular points of information and focuses on them. The basic assumption of ABS is that rational agents with a limited perception act on partial information and concentrate on a task thereby. The information provided by static or dynamic pictures and enriched linguistic and world knowledge with voxemes is too much for these agents to act properly. The annotation focuses on particular viewpoints on objects or aspects of information conveyed by language. Pictures carry too much information, while the annotated language, for instance in dialogues, focuses on a *small part* of it with perspectives. With this annotated partial information with particular views, the agents focus on their task and act intelligently. This paper claims such a focused interaction between the small restricted environment and the task is a fact.

2. Background study

Here are two views of dialogue. I use these views as a background study when analyzing and interpreting dialogues.

2.1. Classical Common Views of Dialogues

Dialogues are interactive linguistic exchanges among at least two participants, conveying or receiving information for actions or emotive reactions. The participants are message senders, recipients, and others directly or indirectly involved with specific intentions or forced responses, differentiating the various types of *dialogue acts* (Bunt, 2019; ISO, 2020). These participants can be either human agents or artificial rational agents like robots.

Question-answering is a typical type of dialogue. One party raises a question, while the other responds if a dialogue succeeds. Negotiations constitute another type of dialogue: one party proposes by requesting, while the other party accepts, modifies, or rejects the proposal by taking linguistic or non-linguistic actions. There may be mediators.

Dialogues are heavily grounded in various types of participant attitudes, background situations, and affordances. They thus license qualifications, restrictions, redundancies, or utterance omissions, much depending on their described situations, as spelled out by Barwise and Perry (1983) and with their later work on situation theory and semantics.

2.2. Dialogues in the Vox World

Pustejovsky and Krishnaswamy (2014, 2016) introduces a modeling language VoxML for visual object conceptual structures in language actions. As stated in Section 1, one of the key notions in VoxML is the *Vox World*. Pustejovsky and Krishnaswamy define this notion more formally with rich implications as a multimodal simulation framework for modeling embodied human-computer interactions and communication between agents engaged in a shared goal or task.

In the Vox World, dialogues are modeled as part of HOI (human-object interactions) or HCI (human-computer interaction) through language. Dialogue participants can be humans (H) or computers (C), all as rational agents that may include artificial agents like computers, while some other objects also participate or get involved indirectly in dialogues. Task-oriented dialogues are embodied interactions between agents, where language, gesture, gaze, and actions are situated within a common ground shared by all agents in the communication. Situated semantic grounding assumes a shared perception of agents with co-attention over objects in a situated context, with co-intention towards a common goal. Dialogues are thus viewed as complex linguistic phenomena in the Vox World.

3. Issues in Interpreting Dialogues

In this section, some dialogues are presented to focus on the issues of interpreting them with illustrations.

3.1. Interpreting Dialogue 1

Dialogue 1 illustrates the complexity of actions even in a short dialogue. It shows how a husband and his wife interact with each other in a shared task of making a cocktail punch and tasting it. The spoken part of the dialogue itself is simple, consisting of three words: "try," "it," and "delicious."

- (1) Dialogue 1
Husband: Try it.
Wife: Delicious!

The script alone cannot be understood unless a situation, depicted visually like Figure 2, is given as a background. The script, as it is, only tells that it is a dialogue between two participants (a husband and his wife) and that the pronoun "it" refers to something edible or potable with a taste. The verb "try" means "try to eat or sip and see how it tastes." It is a task-oriented dialogue through which the couple tries to work together on some common goal, namely to make a good cocktail punch.



Figure 2: Dialogue Situation Visualized

Figure 2 supports the situation in which the dialogue has developed.¹ Two dialogue participants, the husband and his wife, are holding a glass together. The glass looks like containing a cocktail punch. The couple may have been preparing a good punch, possibly for a party. The husband

¹This picture is provided by Ghang Lee (2023), who worked on it through *Dalle3+ChatGpt*.



Figure 3: Empty Glass to be Washed

made a cocktail punch in a punch bowl, poured part into a small container from which one can drink, and handed over the glass to his wife to taste by saying, "Try it." See the difference between a punch bowl and a punch glass: you won't lift the bowl and try the punch from it. Here, the pronoun "it" refers to the punch the husband prepared, but it could have referred to anything edible or that can be tasted. What was presented to the wife was the glass containing the punch. So the wife took the glass of cocktail punch in her hand, raised it to her mouth, sipped the punch, and said "Delicious!". The utterance of the single word "Delicious" followed a series of actions with satisfaction on her facial expression. The wife tasted the punch the husband prepared, and approved the husband for it, making him feel good.

All these actions are not shown in Figure 2. The dialogue implies them only when the picture is looked at. Both the dialogue and the picture are *interpreted* coherently. An adequate interpretation model should be constructed to interpret the cooperative roles of the two dialogue participants, the wife and the husband, who made a punch and tasted it, and the two objects, the glass and the punch contained in it. The punch bowl and other material not shown in the picture may have been somewhere in the kitchen.

3.2. Interpreting Dialogue 2

Dialogue 2 is even shorter than Dialogue 1. It is a short script with two words, supposedly for a

dialogue between a couple, Husband and Wife.

(2) Dialogue 2

Husband: Take this.

Wife: [says nothing.]

Dialogue 2 records the husband uttering the two words "Take this.", asking the wife to take something that is referred to by the demonstrative pronoun "this" and should be located near the speaker himself, but the wife says nothing. There were *two dialogue participants*, and the husband's act was *task-oriented*, telling or ordering his wife to take something near him. This is all that a dialogue act annotation can capture.

A visualized situation, depicted with Figure 3,² for the dialogue provides detailed information on the interactions between the husband and the wife. The wife didn't say a word, but one should see her face in the picture, Figure 3. It says a lot. A husband, sitting on a couch in the living room, told his wife, standing by the dishwasher, to take the glass in his hand, expecting her to put it in the dishwasher. The wife was angry at her husband, who played the king. A situation like this may be considered disgusting in some cultures.

3.3. Dialogue 3 in Contrast to Dialogue 2

A dialogue almost the same as Dialogue 2 has a totally different interpretation. In Dialogue 3, the wife expresses her appreciation.

²Ghang Lee also provided this image.

- (3) Dialogue 3 with Appreciation
 Husband: Take this!
 Wife: Thanks. Looks delicious.

The husband mixed a cocktail punch and *offered* it to his wife. The wife says "Thanks" in an appreciative way by saying a little more, "Looks delicious." The following picture³ depicts a delightful scene that says more than words.



Figure 4: Punch offered to the Wife

I have presented the two pictures that visualized dialogue situations. They show how much visual information contributes to the rich interpretation of dialogues or interactive communications. The same imperative "Take this!" is interpreted differently, one as an *order* and the other as an *offer*.

3.4. Dialogue 2 Extended

Dialogue 4 illustrates with Script 4 how Dialogue 2 is extended with another round of exchanging the turns.

- (4) Dialogue 4 Extending Dialogue 2
 Husband: Take this.
 Wife: [Got angry, saying nothing.]
 Husband: Sorry. I'll do it.
 Wife: [Facial expression changed to exasperation. She is still silent.]

Looking at his wife's angry face, the husband realized he had mistakenly asked her to take the glass to the dishwasher. He thus apologized and took the glass himself to the washer.

The dialogue has four turns, although the wife does not respond verbally. Such a situation can easily be imagined and turned into a short video.

³Ghang Lee also provides this image.

However, the current technology has not fully developed to convert text to videos.⁴

As one of the reviewers pointed out, it must be emphasized that it is not so much the picture itself providing the background context of the dialogue but rather the situation type we construct based on the picture. We imagine or visualize appropriate situations or create such scenes to interpret dialogues. Dialogues, on the other hand, help interpret visually perceptible scenes by helping us focus on some specific parts of them.

4. Annotating Dialogues

4.1. Basic Annotation Structure of Dialogues

The annotation of dialogues follows Bunt (2019) and ISO (2020). The basic structure of a dialogue consists of two parts, the dialogue act and the semantic content. In the simplest case, the dialogue structure is a quadruple $\langle\langle s, A, f_d \rangle, c\rangle$, where the triple represents the simplest dialogue act structure consisting of a sender s , addressees A , and a dimension-specific function f_d while the last component c represents the dialogue content. For general purposes, this list can be extended to the most complex case with a 7-tuple (ISO, 2020) plus the content c , where the three bracketed components need not be specified:

- (5) $\langle\langle s, A, [h], f, d, [q], [E] \rangle, c\rangle$ of attributes,
 where s is a sender (speaker),
 A addressees,
 H other participants,
 f a general-purpose communicative function,
 d a dimension,
 q qualifiers,
 E dialogue units that the act depends on, and
 c the semantic content of the dialogue.

The first seven components specify the act type of dialogues while the last component c refers to the dialogue content. The content c directly *plugs in* the semantic content, which carries the information of a dialogue associated with a dialogue act. The 7-tuple plus the content c forms a complex feature structure such that the value of c is directly linked to another annotation scheme. No link like `contentLink` needs to be introduced, although it is a preference recommended in Bunt (2019) and proposed in ISO (2020).

4.2. VoxML-linked Annotation

The VoxML-linked annotation (Lee et al., 2023) refers to the Voxicon, a component of VoxML, consisting of complex feature structures, called *vox-*

⁴The Open AI just announced Sora for such a task.

emes. These voxemes represent the visual object conceptual structures of VoxML basic categories such as **object**, **program** (event, motion, or action), and **relation** that includes property and function). Each voxeme is associated with a linguistic expression (e.g., "glass"), its morpho-syntactic or lexical information, and semantico-pragmatic or physical information associated with it such as information about its habitat, affordance structures, and embodied interactions (Pustejovsky, 1995; Gibson, 1977; Pustejovsky and Krishnaswamy, 2016, 2021). The reference to these structures is expected to free the VoxML-oriented ABS from its reliance on syntactic or pragmatic analysis.

For illustration, consider the annotation of Dialogue 3. The annotation takes two steps: Step 1 focuses on the dialect act, while Step 2 on its content.

4.2.1. Step 1: Annotating Dialogue Act

The first part of the whole script, which includes the information about the speaker and the addressee, is annotated as in (6).

- (6) Annotating the Dialogue Act of Dialogue 3
- a. Segmented Dialogue Script (id="d3S"): Husband_{w1} to Wife_{w3}: Take_{w4} this_{w5}.
 - b. Dialogue Act Annotation:


```
<dialogue id="#d3", target="#d3S">
  <dAct id="d3A", sender="#w1",
    addressee="#w3", dimension="task",
    cFunction="offer", content="#d3C"/>
</dialogue>
```

The dialogue act annotation marks up not just what has been uttered by the speaker, but the whole dialogue script that describes all the components that constitute the act of a given dialogue.

4.2.2. Step 2: Annotating the Content

The proposed VoxML-linked ABS annotates the content c of a dialogue by referring to the dialogue utterance and the background situation, possibly depicted by an associated picture. The content of Dialogue 3 is annotated as in (7):

- (7) Annotating the Content of Dialogue 3
- ```
<dialogue id="#d3", target="#d3S">
 <dContent id="d3C", linkedTo="#d3A">
 <object id="o1", target="#w1"
 type="human", pred="husband",
 relatedTo="#w3"/>
 <object id="o2", target="#w3"
 type="human", pred="wife",
 relatedTo="#w1"/>
 <action id="a3", target="#w4"
 type="transition",
```

```
 pred="take:consume"5,
 agent="#o2", theme="#o6:punch"/>
 <object id="o3", target=" ",
 type="physicalObj:artifact",
 pred="glass", definite="yes",
 grabbedBy="#o4:hand",
 comment="See Figure 4"/>
 <object id="o4", target="",
 type="physicalObj", pred="hand",
 definite="yes", partOf="#o1:husband",
 comment="See Figure 4"/>
 <object id="o5", target="",
 type="physicalObj:liquid:beverage",
 pred="punch", definite="yes",
 containedIn="#o3:glass",
 comment="See Figure 4" />
</dContent>
</dialogue>
```

With the comment "See Figure 4", the demonstrative pronoun "this" is annotated as referring to the punch in the glass held by the husband in his hand. It does not refer to the glass, for it is already in the wife's hand. The verb "take" is thus understood as meaning *to consume the punch*, instead of meaning *to grab the glass with a hand*. The dialogue does not mention "glass," "punch," or "hand" but the annotation introduces them all as *non-consuming tags*. Figure 4 shows that the glass is in the husband's hand and also in the wife's hand.

#### 4.3. Abstract Syntax and the Metamodel

The annotation of the content structure in the Vox World as presented in (7) requires the specification of an annotation scheme. Such a specification is done partially with the formulation of an abstract syntax. For this, the abstract syntax, named  $ASyn_{vox}$ , is minimally formulated for the annotation in the Vox World, as in (8):

- (8)  $ASyn_{vox}$  is defined as a tuple  $\langle M, B, @ \rangle$ , such that, given a language  $L$ ,
- a.  $M$  is a nonempty subset, called *markables*, of  $L$ , delimited by  $B$ ;
  - b.  $B$  is a set  $\{o, a, r\}$  of *base categories*:
    - $o$  stands for category **object**;
    - $a$ , category **action**, a subcategory of **eventuality**;
    - $r$ , category **relation** that includes the subcategories **function** and **property**;
  - c.  $@$  is a set of *assignment functions* from features (attributes) to values associated with each category in  $B$ .

Note that this syntax has no links. Instead, some attributes are plugged into other annotation structures. See Annotation Structure (9).

<sup>5</sup>See WordNet-3 for the sense of "take."

- (9) Semantic Roles:  
`<action id="a3" type="transition", pred="take", agent="#o2", theme="#o3"/>`

The semantic roles for the action *take* are directly annotated into its base annotation structure by referring to the semantic role frames in a lexicon. There is no repeated application of a link like `srLink` for semantic role labeling.

The minimal abstract syntax  $\mathcal{ASyn}_{vox}$  specified in (8) conforms to the metamodel for the Vox World as a markup language (Lee et al., 2023).

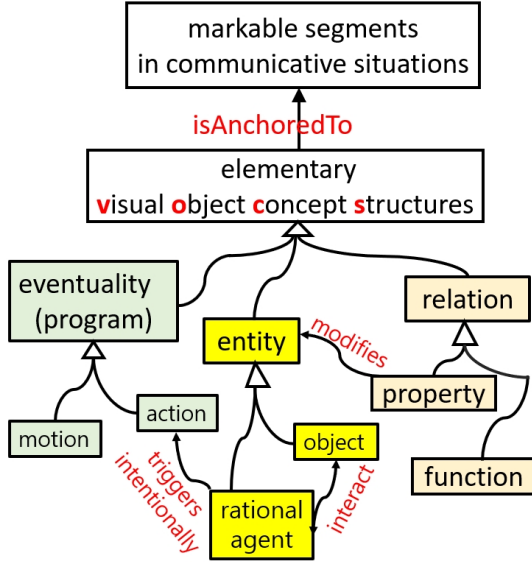


Figure 5: Metamodel of the Abstract Syntax

Here, the Vox World consists of anything perceptible in communicative situations including dialogues. Visual object concept structures (vocs or vox), either elementary or their relational compositions, are then anchored to the Vox World. The vox are categorized into three major categories with subcategories **eventuality (program):action**, **object:rational agent**, and **relation:property, function**. Actions are triggered by rational agents intentionally, while rational agents, either humans or robots, interact with one another or other objects. Properties (attributes) modify objects, while functions and relations operate or range over visual object conceptual structures.

## 5. Interpretation

### 5.1. Overview

ABS (Lee, 2023) interprets annotation structures for a model constrained by relevant parts of the Vox World. Implementing classical set-theoretic models as in Montague semantics (Montague, 1974b), or the Discourse Representation Theory (DRT) (Kamp and Reyle, 1993; Parson, 1990), these parts of the Vox world supplement those models  $\langle D, R, [[ ]] \rangle$

of denotational semantics, especially by formally delimiting the domain  $D$  of a model, which normally consists of individual entities, and the set of  $n$ -ary relations  $R$  over  $D$  or its Cartesian products with a small world in which some relevant visual object concept structures reside.

In the Vox World, everything in its small world is a first-class citizen, including properties and relations, as in Situation Semantics (Barwise, 1989), or else the notion of *functional types* is introduced to allow such objects as *event descriptors* of type  $e \rightarrow t$  (Kracht, 2002; Pustejovsky et al., 2019) or as in Davidsonian Semantics (Davidson, 1967, 2001; Parson, 1990). ABS then interprets annotation structures in two steps. First, annotation structures  $\mathbf{a}$  are translated to semantic forms  $\sigma(\mathbf{a})$  in typed first-order logic. Second, these logical forms are interpreted for a well-defined model  $M$  constrained by the Vox World  $v$ :  $[[\sigma(\mathbf{a})]]^{M,v}$ .

### 5.2. Translating Annotation Structures to Logical Forms

To interpret annotation structures, ABS translates them into semantic forms directly. ABS does not require syntactic analysis to derive semantic forms because the annotation already contains the necessary information for adequate translation. In contrast, Montague Semantics (Montague, 1974b) uses Categorial Grammar for analyzing input data to trees, for instance, to capture scope ambiguity, before translating the analyzed trees to semantic forms in Higher-order Intensional Logic.

Translation (10) shows how the annotation structures of category **object** are translated.

- (10) a. `<object id="o1", target="#w1" type="human", pred="husband", relatedTo="#w3"/>`  
 $\sigma(o1) := [human(x_1), husband(x_1, x_2)]$   
 b. `<object id="o2", target="#w3" type="human", pred="wife", relatedTo="#w1"/>`  
 $\sigma(o2) := [human(x_2), wife(x_2, x_1)]$

The attribute `@relatedTo` in the annotation structures treats the predicates *husband* and *wife* as binary relations in the semantic forms.

The transitive verb "take" denotes an action of type *transition* with two required arguments. Translating the annotation structure that marks up its semantic content is straightforward. The two semantic roles associated with the two verb arguments are marked up.

- (11) `<action id="a3", target="#w4", type="transition", pred="take", agent="#o2", theme="#o3"/>`  
 $\sigma(a3) :=$

$[transition(e_3), take(e_3), agent(e_3, x_2), theme(e_3, x_3)]$

The semantic form  $\sigma(a_3)$  here does not add new information to the annotation. The predicate *take* is a transition, thus involving a series of sub-actions: the wife, who was told to take something, referred to with the demonstrative pronoun "it", must reach a reachable position to grab the object and take it out, intending to move it to somewhere for some purpose. The annotation does not capture such information but must be captured at the interpretation stage, given an appropriate background.

### 5.3. Direct Interpretation vs. Enriched Logical Forms

Intuitively speaking, annotation structures should be interpretable without being translated into logical forms, as in Montague (1974a)'s English as a Formal Language. Translation carries no additional meaning except that it shows that the translated logical forms are expressed in lower-order logic. However, it is possible to generate enriched annotation structures by referring to the Voxicon.

VoxML contains the Voxicon that lists *voxemes* enriching annotation structures of those categories, **object**, **event**: **action**, and **relation**: **property**, **function** in the metamodel. For illustration, consider the annotation structure of category **object**: `<object id="o3", target="#o4 (glass)"/>` to enrich it with the voxeme of *glass* listed in the Voxicon.

|              |                                                                                                                                                                                                                                                                                                                           |
|--------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>glass</b> |                                                                                                                                                                                                                                                                                                                           |
| LEX =        | $\left[ \begin{array}{l} \text{PRED} = \text{glass} \\ \text{TYPE} = \text{physobj, artifact} \end{array} \right]$                                                                                                                                                                                                        |
| TYPE =       | $\left[ \begin{array}{l} \text{HEAD} = \text{cylindroid}[1] \\ \text{COMPONENTS} = \text{surface, interior} \\ \text{CONCAVITY} = \text{concave} \\ \text{ROTATSYM} = \{Y\} \\ \text{REFLECTSYM} = \{XY, YZ\} \end{array} \right]$                                                                                        |
| HABITAT =    | $\left[ \begin{array}{l} \text{INTR} = [2] \left[ \begin{array}{l} \text{CONSTR} = \{Y > X, Y > Z\} \\ \text{UP} = \text{align}(Y, \mathcal{E}_Y) \\ \text{TOP} = \text{top}(+Y) \end{array} \right] \\ \text{EXTR} = [3] \left[ \text{UP} = \text{align}(Y, \mathcal{E}_{\perp Y}) \right] \end{array} \right]$          |
| AFFORD_STR = | $\left[ \begin{array}{l} A_1 = H_{[2]} \rightarrow [\text{put}(x, \text{on}([1]))] \text{support}([1], x) \\ A_2 = H_{[2]} \rightarrow [\text{put}(x, \text{in}([1]))] \text{contain}([1], x) \\ A_3 = H_{[2]} \rightarrow [\text{grasp}(x, [1])] \\ A_4 = H_{[3]} \rightarrow [\text{roll}(x, [1])] \end{array} \right]$ |
| EMBODIMENT = | $\left[ \begin{array}{l} \text{SCALE} = \text{<agent} \\ \text{MOVABLE} = \text{true} \end{array} \right]$                                                                                                                                                                                                                |

Figure 6: Voxeme of a Glass

The voxeme of *glass* in Figure 6 represents five sorts of information: [i] LEX, [ii] TYPE, [iii] HABITAT, [iv]

AFFORD\_STR, [v] EMBODIMENT.<sup>6</sup> Annotation (12b) represents part of the lexical information (LEX) and the affordance structure (AFFORD\_STR) information about its being a container ( $A_2$ ). The EMBODIMENT says that the object's size is smaller than the agent who carries or grabs it and can be moved by the agent. For illustration, consider annotating the noun "glass" in Dialogue 3. Its annotation structure can be enriched with contextual information by referring to the voxeme as in Figure 6.

(12) a. Basic Annotation, copied from 7:

```
<object id="o3", target=" ",
type="physicalObj:artifact",
pred="glass", definite="yes",
grabbedBy="{#o1,#o4:hand}",
refersTo="Figure 4"/>
```

b. Annotation Enriched with Voxeme 6:

```
<object id="o3", target=" ",
type="physObj:artifact",
pred="glass", definite="yes",
grabbedBy="{#o1,#o4:hand}",
form="cylindroid", shape="concave",
use="container", contains="o5:punch",
smallerThan="{#o1,#o4}",
refersTo="Figure 4, Figure 6"/>
```

c. Logical Form  $\sigma(x_3) :=$

```
 $[physobj(x_3), artifact(x_3),$
 $glass(x_3), definite(x_3),$
 $grab(e_1), agent(e_1, x_1), theme(e_1, x_3),$
 $instrument(e_1, x_4 : husband'sHand),$
 $cylindroid(x_3), concave(x_3),$
 $container(x_3), contains(e_2),$
 $theme(e_2, x_6 : punch)]$
```

Annotation (12b) shows the enrichment of Annotation with some pieces of information obtained from the voxeme of "glass" presented in Figure 6. The logical form based on the enriched annotation states that the glass, which is small enough to be grabbed by the husband, contains punch.

### 5.4. Interpretation

The Vox World provides visual information for interpreting actions and interactive communications. Specifically, it controls the three processes of annotating, translating, and interpreting dialogue acts and contents interchanged among the participants. The utterance "Take this.", which is made by the husband in the two different dialogues, for instance, is annotated differently: in Dialogue 2, it is annotated as an *order*, whereas it is annotated as an *offers* in Dialogue 3. In Dialogue 2, the demonstrative pronoun "this" refers to the empty glass. In

<sup>6</sup>For the detailed explanation of the voxeme of *glass*, see Lee et al. (2023).

Dialogue 3, in contrast, the same pronoun refers to either the glass with a punch in it or the punch in the glass, for the wife says, "Looks delicious," referring to the punch, not the glass.

Annotation and the Vox World complement each other. Voxemes enrich annotation structures. Annotation can capture all these differences and refer to the appropriate figures for appropriate information, but the voxemes alone cannot.

Annotation structures and semantic forms are inadequate to capture finer-grained information associated with all aspects of dialogues. This especially concerns the interpretation of actions, for actions of type transition particularly involve a dynamic sequence of sub-events or sub-actions. The husband's order in Dialogue 2 is not a simple act, but a complex sequence of sub-situations and sub-actions.

- (13) Sub-situations and sub-actions in Figure 2:
- a. The wife was standing near the washing machine.
  - b. The husband was sitting on a sofa not far from the kitchen.
  - c. The husband asked the wife to take the glass,
  - d. and expecting
  - e. her to come to him easily
  - f. to pick it up from his hand and
  - g. put it in the dishwasher.
  - h. Her emotional reaction, displayed on her face with silence,
  - i. indicated that his expectation was wrong.
  - j. She rather expected
  - k. him to come and
  - l. put the glass in the dishwasher himself.

All this information cannot be captured in the annotation or represented in simple logical forms. It can only be *abduced*<sup>7</sup> by learning relevant perspectives on the informational content and the intention of dialogue or discourse participants, as mentioned by Hobbs (1996). In addition, such an abduction becomes possible by constructing appropriate background scenes with visual object conceptual structures (vox). The construction of such scenes is systematically constrained in the Vox World that characterizes not only the lexical features of the language used in human communications, but also the habitat, affordance structures, and embodiment of objects and actions, and their interactions mentioned in that language with perceptual (visual) conditions.

---

<sup>7</sup>I have intentionally used the term *abduce* to focus on the experiential and perceptual aspects of Peirce (1931–1958); Hobbs et al. (1993); Hobbs (1996, 2006) for understanding language and logic.

## 6. Concluding Remarks

The partiality of information is a basic motivation for annotation, for annotation marks up only some parts of a language. This paper has shown how this notion of partiality works in annotating and interpreting dialogues. Annotation also explicitly uses language such as dialogues by annotating the type of dialogue acts and content and interpreting them against a small visual world called the Vox World.

The paper treated the tripartite understanding of dialogues: annotation, visualization, and interpretation. Annotation focuses on some basic linguistic elements in described situations in which dialogue participants interact with relevant objects or each other. At the same time, visualization provides details of fine-grained perspectives with background information. Interpretation with logical forms validates such details of information with consistency.

The paper proposes using visual information in general and the Vox World in particular, to annotate and interpret dialogues or other interactive communications among rational agents or relevant objects. It even suggested that a set-theoretic semantics should be redesigned by restructuring its basic model structure  $\langle D, R, [[]] \rangle$ . For instance, the domain  $D$  and the set  $R$  of  $n$ -ary relations can be modified with a small set of visual object concept structures. Or else, such a model is minimally implemented but constrained by something like the Vox World. However, the formal specification of such a task is left for the future.

The paper intentionally focused on simple dialogues to highlight the complementary roles of dialogue scripts and related images and on the role of VoxML-linked annotation that links them for coherent interpretation. Complex dialogues, such as those involving misunderstandings and subsequent repair strategies, require complex images, such as motion pictures, for their interpretation.

Pictures are extensively used to show how dialogues are annotated and interpreted. For this reason, the proposed VoxML-linked ABS may be understood mistakenly as a picture-based semantics that requires the generation of static or dynamic pictures as an essential process. It is a total misunderstanding. Pictures help visualize the situations in which dialogues are possibly enacted. Humans can easily visualize such situations through the power of imagination. It is, however, a different question of how artificial agents learn to visualize dialogue situations and interpret them or even to participate in a dialogue by understanding the flow of dialogues. Such a question is left for future work.

## 7. Acknowledgments

I owe many thanks to the three anonymous reviewers who provided detailed constructive comments to improve the paper extensively and to Jae-Woong Choe, Minhaeng Lee, and Chong-won Park, who helped write the preliminary version of the paper. I also would like to thank Byonrae Ryu, who reset the figures, and Ghang Lee with the production of dialogue-related images for their time-consuming work.

## 8. Bibliographical References

- Jon Barwise. 1989. *The Situation in Logic*. CSLI (Center for the Study of Language and Information, Stanford, CA).
- Jon Barwise and John Perry. 1983. *Situations and Attitudes*. The MIT Press, Cambridge, MA.
- Harry Bunt. 2019. Plug-ins for content annotation of dialogue acts annotation. In *Proceedings of the 15th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-15)*, pages 33–45, Workshop at the 11th International Conference on Computational Semantics (IWCS 2019), Gothenburg, Sweden, May 23, 2019.
- Harry Bunt. 2022. Intuitive and formal transparency in semantic annotation schemes. In *Proceedings of the 18th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-18)*, pages 102–109, Workshop at LREC2022, Marseilles, France.
- Donald Davidson. 1967. The logical form of action sentences. In N. Rescher, editor, *The Logic of Action and Decision*, pages 81–120. University of Pittsburgh Press, Pittsburgh, PA.
- Donald Davidson. 2001. *Essays on Actions and Events*. Oxford University Press, Oxford.
- David Dowty, Stanley Peters, and Robert Wall. 1981. *Introduction to Montague Semantics*. Reidel, Dordrecht.
- James Jerome Gibson. 1977. The theory of affordances. *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pages 67–82. Reprinted as chapter 8 of Gibson (1979).
- James Jerome Gibson. 1979. *Ecological Approach to Visual Perception*. Psychology Press, New York.
- Jerry R. Hobbs. 1996. On the relation between the informational and intentional perspectives on discourse. In E. Hovy and D. Scott, editors, *Computational and Conversational Discourse: Burning Issues— An Interdisciplinary Account*, pages 247–260. Springer, Berlin, Germany.
- Jerry R. Hobbs. 2006. Abduction in natural language understanding. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*, chapter 32, pages 724–741. Blackward Publishing, London. <https://doi.org/10.1002/9780470756959.ch32>.
- Jerry R. Hobbs, E. Stickel, Mark, Douglass Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63(1-2):69–142.
- ISO. 2020. *ISO 24617-2 Language resource management – Semantic annotation framework – Part 2: Dialogue acts*. International Organization for Standardization, Geneva. 2nd edition.
- Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, Dordrecht.
- Marcus Kracht. 2002. On the semantics of locatives. *Linguistics and Philosophy*, 25:157–232.
- Kiyong Lee. 2020. Annotation-based semantics. In *Proceedings of the 16th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-16)*, pages 37–49.
- Kiyong Lee. 2023. *Annotation-Based Semantics for Space and Time in Language*. Cambridge University Press, Cambridge, UK.
- Kiyong Lee, James Pustejovsky, and Nikhil Krishnaswamy. 2023. An abstract specification of VoxML as an annotation language. In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-19)*, pages 66–74, June 20, 2023, Workshop at IWCS 2023, Nancy, France. ACL anthology L16-1730.
- Richard Montague. 1974a. English as a formal language. In *Formal Philosophy: Selected Papers of Richard Montague*, New Haven and London. Yale University Press.
- Richard Montague. 1974b. The proper treatment of quantification in ordinary english. In *Formal Philosophy: Selected Papers of Richard Montague*, New Haven and London. Yale University Press.
- Terence Parson. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. The MIT Press, Cambridge, MA.

- Charles S. Peirce. 1931–1958. *Collected Papers of Charles Sanders Peirce*. Harvard University Press, Cambridge, MA. Edited by Hartshorne, C. and Weiss, P. and Burks, A.
- James Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, MA.
- James Pustejovsky. 2013. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10. Association for Computational Linguistics, Pisa, Italy.
- James Pustejovsky and Olga Batiukova. 2019. *The Lexicon*. Cambridge University Press.
- James Pustejovsky and Nikhil Krishnaswamy. 2014. Generating simulations of motion events from verbal descriptions. In *Proceeding of the 3rd Joint Conference on Lexical and Computational Semantics. (\*SEM 2014)*, pages 99–109.
- James Pustejovsky and Nikhil Krishnaswamy. 2016. VoxML: A visualization modeling language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4606–4613, Portorož, Slovenia. ELRA. ACL anthology L16-1730.
- James Pustejovsky and Nikhil Krishnaswamy. 2021. Embodied human-computer interaction. *KI-Künstliche Intelligenz*, 35(3-4):307–327.
- James Pustejovsky, Kiyong Lee, and Harry Bunt. 2019. The semantics of ISO-Space. In *Proceedings of the 15th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-15)*, pages 46–53. May 23, 2019, at IWCS2019, Gothenburg, Sweden.

## 9. Copyrights

The Language Resources and Evaluation Conference (LREC) Proceedings are published by the European Language Resources Association (ELRA). They are available online from the conference website.

ELRA's policy is to acquire copyright for all LREC contributions. In assigning your copyright, you are not forfeiting your right to use your contribution elsewhere. This you may do without seeking permission and is subject only to normal acknowledgment to the LREC proceedings. The LREC Proceedings are licensed under CC-BY-NC, the Creative Commons Attribution-Non-Commercial 4.0 International License.



# Annotating Evaluative Language: Challenges and Solutions in Applying Appraisal Theory

**Jiamei Zeng**

Department of Linguistics and  
Translation  
City University of Hong Kong  
Hong Kong SAR  
jjamezeng3-c@my.cityu.edu.hk

**Min Dong**

School of Foreign Languages  
Beihang University  
PR China  
mdong@buaa.edu.cn

**Alex Chengyu Fang**

Department of Linguistics and  
Translation  
City University of Hong Kong  
Hong Kong SAR  
acfang@cityu.edu.hk

## Abstract

This article describes a corpus-based experiment to identify the challenges and solutions in the annotation of evaluative language according to the scheme defined in Appraisal Theory (Martin and White, 2005). Originating from systemic functional linguistics, Appraisal Theory provides a robust framework for the analysis of linguistic expressions of evaluation, stance, and interpersonal relationships. Despite its theoretical richness, the practical application of Appraisal Theory in text annotation presents significant challenges, chiefly due to the intricacies of identifying and classifying evaluative expressions within its sub-system of Attitude, which comprises Affect, Judgement, and Appreciation. This study examines these challenges through the annotation of a corpus of editorials related to the Russian-Ukraine conflict and aims to offer practical solutions to enhance the transparency and consistency of the annotation. By refining the annotation process and addressing the subjective nature in the identification and classification of evaluative language, this work represents some timely effort in the annotation of pragmatic knowledge in language resources.

**Keywords:** Appraisal Theory, Attitude, evaluative language, pragmatic annotation

## 1. Introduction

Appraisal Theory (Martin and White, 2005) describes a taxonomy of semantic resources that allow for the expression of emotions, judgements, and valuations as well as the means to enhance and engage with these evaluations (Martin 2000, p.145). It has attracted an increasing academic interest evidenced by a growing volume of publications in the Web of Science (Figure 1), indicating the urgent need for the pragmatic analysis of evaluative language.

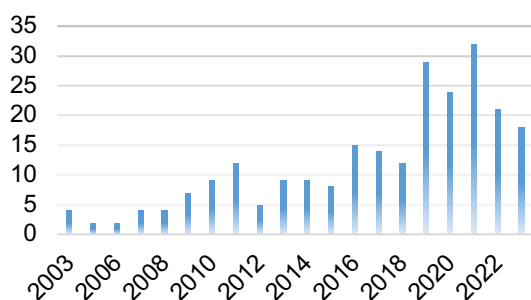


Figure 1: Annual count of academic publications on Appraisal Theory from 2003 to 2023

Considered as a systematic, detailed and elaborate framework for the analysis of evaluative language (Bednarek, 2006, p. 32), Appraisal Theory has demonstrated a great expanding relevance across various fields including, among many others, the examination of academic discourse (e.g. Swain, 2010; Hood, 2010; Geng and Wharton, 2016), political language (e.g. Mayo and Taboada 2017), news narratives (e.g.

Bednarek and Caple, 2010; Huan, 2016), business discourse (e.g. Pounds, 2011; Fuoli and Hommerberg, 2015), wine tasting sheets (Breit, 2014), movie reviews (Taboada et al., 2014), and public statements (Meadows and Sayer, 2013).

However, as a sophisticated analytical framework involving semantic and pragmatic interpretations, the theory is not without its challenges, particularly when applied to the annotation of large corpora of natural texts. A major challenge lies in the dual tasks of annotation practices: identifying textual elements of appraisal and classifying them according to the theory's component categories of Attitude, Engagement, and Graduation and their respective sub-categories (Fuoli, 2018). This complexity is compounded by the inherent subjectivity and variability of linguistic expressions.

Fuoli (2018) suggests a step-wise method as a general solution. Our work to be reported next aims to provide more detailed solutions by targeting the Attitude category and addressing specific problems and issues, thereby exploring the issue of operationality through clear, operable strategies. In particular, we constructed a corpus of editorials from news media, performed the annotation of this material according to the Attitude system, and reviewed the various problematic issues before the formulation of solutions. We aim to offer additional insight about the aspects of applying a theoretically rich but operationally challenging framework through practical annotation of a sound level of transparency and consistency. We also hope that efforts such as ours will help to harness the full

potential of Appraisal Theory for the analysis and understanding of evaluative language.

## 2. Methodological Issues

This section provides a comprehensive outline of the methodological framework applied in the appraisal annotation of editorial content. We will first explain the rationale behind the selection of editorials as the primary material and introduce the composition of the annotator team. Following this, an in-depth examination of the chosen annotation framework, the tool utilized, and the procedural steps undertaken will be presented. These elements collectively form the foundation of our systematic approach.

### 2.1 Corpus Data

A corpus was constructed comprising editorials, selected for their inherent nature of presenting opinions, making them an ideal subject for this study. Four diverse newspapers were selected as the primary sources of data, including China Daily (CD), New York Times (NYT), South China Morning Post (SCMP), and The Guardian (TG). Thirty editorials were selected from each newspaper, all of which were published between January 2022 and May 2023 and centred on the Russian-Ukraine conflict, amounting to a total of 120 articles. The corpus of editorials is summarized in Table 1. This time frame and subject matter were set up to capture a wide range of evaluative perspectives during a period of significant geopolitical tension.

|       | CD     | NYT    | SCMP   | TG     | Total  |
|-------|--------|--------|--------|--------|--------|
| Text  | 30     | 30     | 30     | 30     | 120    |
| Token | 14,073 | 15,170 | 20,975 | 18,551 | 68,769 |
| Type  | 2,982  | 3,368  | 4,255  | 4,035  | 8,678  |

Table 1: Summary of the corpus of editorials

### 2.2 Annotation Framework

Appraisal System is defined as the linguistic mechanisms through which authors or speakers express their positive or negative assessments regarding the subjects, events, and situations discussed in their texts (Martin and White, 2005, p. 2). It is divided into three primary systems: Attitude, Engagement, and Graduation, each with its own sub-systems or categories. Our annotation experiment focused on the Attitude system, which comprises Affect (emotional responses), Judgement (evaluations of human behaviour and character), and Appreciation (assessments of objects, texts, events, and processes). Each dimension features a polarity aspect, allowing classifications as either positive or negative.

Affect is the core sub-system of Attitude and is subdivided into four categories: Dis/inclination, Un/happiness, In/security, and Dis/satisfaction. Judgement is divided into two sub-systems including social esteem and social sanction.

Social esteem relates to the evaluation of someone's abilities (Capacity), their adherence to norms (Normality), and their persistence or determination (Tenacity). Social sanction focuses on truthfulness (Veracity) and appropriateness or morality (Propriety). Appreciation evaluates reactions to, compositions of, and valuations of objects or phenomena.

### 2.3 Annotators and Annotation Tool

The annotation of the corpus was performed by six MA students in linguistics, divided into three annotation groups with two annotators each. UAM Corpus Tool (O'Donnell, 2008) was chosen as the annotation tool for the experiment. It has a user-friendly interface and provides modules for statistical analysis of the annotated data. This feature was useful for the presentation and interpretation of our annotation results.

### 2.4 Annotation Process

The annotation process involved the initial training of the annotators to ensure a sound level of consistency measured in terms of inter-annotator agreement before the full-scale annotation of the corpus was rolled out. The process involved the following specific steps:

Step 1: Each group were first of all required to familiarize themselves with Martin and White (2005) in general and Attitude in particular during the first stage.

Step 2: A tutorial session was given to all the annotators, key concepts summarized and major principles outlined. An annotation guide was drawn up.

Step 3: A first trial annotation was performed simultaneously by the three pairs of annotators on one text (Editorial CD 232323), which consists of 472 tokens. The initial inter-annotator agreement score was extremely low for this task at only 0.267, revealing a broad gap in agreement among the annotators, evidencing the high level of diversity that is expected for the pragmatic annotation of evaluative language.

Step 4: A second training session was carried out. The three annotation groups reviewed relevant aspects of Attitude and discussed the disagreements and problematic issues encountered during the annotation process. This training process eventually resulted in the formulation of a refined set of annotation guidelines.

Step 5: A second trial annotation was conducted on another text of 586 words (Editorial TG 20230223). The annotation this time resulted in a Fleiss kappa score of 0.812, demonstrating a significantly improved and satisfactory level of agreement.

Step 6: The groups proceeded to annotate the remaining corpus independently. The corpus was imported into the UAM Corpus Tool. Although the UAM Corpus Tool comes with some built-in layers for Appraisal Theory, we found it necessary to modify these layers to align with our specific annotation requirements. The resulting layers of the annotation scheme is illustrated in Figure 2. Text segments expressing emotional attitudes were manually identified and marked up through the selection of an appropriate tag.

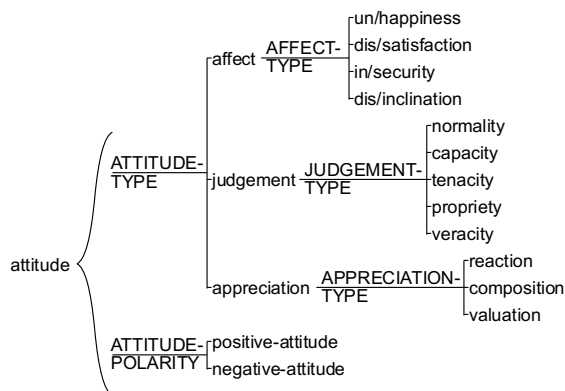


Figure 2: Refined annotation scheme of Attitude

### 3. Principles of Annotation

In what follows, we detail some of the major principles of annotation based on the two tests and outline the specific areas of disagreement encountered during the annotation process. We focus on the identification and categorization of evaluative language in a particular stretch of discourse, aiming to illustrate our practical methodological strategy to capture and classify evaluative expressions within texts with a good level of transparency and consistency.

#### 3.1 Identifying What Needs to Be Annotated

The fundamental step in annotating evaluative language involves discerning which segments of text require annotation. Our principle is to identify and mark the smallest text segment that conveys the overall attitude or evaluative stance, which ensures precision and relevance in our annotations while capturing the attitude embedded within the text. Efforts were made to maintain a full phrase structure. Consider

(1-1) *Wang Huiyao says Beijing is best (+Valuation) placed to help negotiate an end to Russia's war in Ukraine.*

(1-2) *Wang Huiyao says Beijing is best placed (+Capacity) to help negotiate an end to Russia's war in Ukraine.*

In (1-1) and (1-2), we encounter possible annotation segments of “best” and “is best placed”. While “best” alone might suggest a positive Valuation, annotating the broader phrase *is best*

*placed* captures a more specific and contextually rich expression of positive Capacity.

#### 3.2 Contextual Considerations in Annotation

The second principle extends beyond the identification of the smallest meaningful unit to encompass the contextual considerations of nouns that inherently express attitudes. Nouns such as “sanction”, “conflict”, and “invasion”, while potentially evaluative, are approached with caution in specific contexts where they often serve a descriptive role, reflecting the factual dimensions of the situation rather than an evaluative stance. This principle acknowledges the importance of context in determining the evaluative nature of nouns.

#### 3.3 Determining the Specific Category for Annotation

In categorizing annotated items, our approach is informed by principles outlined by Martin and White (2005) and further emphasized by Bednarek (2009). We aimed to differentiate between types of attitudinal lexis and evaluated targets or types of assessment. In practice, this means categorizing expressions related to emotions or feelings of people as Affect, evaluations of behaviour as Judgement, and assessments of objects or phenomena as Appreciation. This classification is instrumental in aligning evaluative expressions with the appropriate domain of appraisal, ensuring that our analysis is both systematic and aligned with the theoretical underpinnings of Appraisal Theory.

Once the primary category is determined, the next step involves specifying the subcategory based on the meaning. This process requires a careful analysis of the text to discern the specific nature of the evaluative stance being expressed. Our principle here emphasizes the importance of a detailed and context-sensitive approach to annotation. Bednarek’s (2009) emphasis on the distinction between types of attitudinal lexis and evaluation targets serves as a crucial reminder of the depth and specificity required in annotating evaluative language, thereby enhancing the analytical precision.

In short, to ensure clarity and consistency, the following principles were applied: identifying the minimal meaningful textual segments for annotation, considering the context to accurately capture evaluative meanings, and categorizing annotations based on types of attitudinal lexis and evaluation targets.

### 4. Problems and Solutions for Annotating Appraisal

In the actual process of annotating evaluative language, despite having established a set of guiding principles, we still encountered several

problems related to identifying and classifying evaluative expressions. This situation underscores the gap between theories and practice, revealing areas that demand refinement, hence suggesting the importance of putting semantic annotation schemes to tests with authentic texts. This section outlines these problems and describes solutions.

#### 4.1 Challenges in Identifying Appraisal and Possible Solutions

In (2) below, the phrase *seeks to* could be interpreted as expressing an inclination, a positive evaluative stance towards the action that follows.

(2-1) *This targeting of civilians reveals that Putin seeks not only to win. He **seeks to** (+Inclination) **demoralize** (-Propriety).*

(2-2) *This targeting of civilians reveals that Putin seeks not only to win. He **seeks to demoralize** (-Propriety).*

However, the verb *demoralize*, which carries a negative connotation (negative Propriety), is the focal point of the evaluative stance in this context. The challenge here concerns whether to annotate *seeks to* for its positive inclination towards an action or to focus solely on the negative evaluative stance conveyed by *demoralize*. To address this issue, we opted not to annotate *seeks to* based on the principle that we should focus on primary evaluative meaning and avoid polarity conflicts. By prioritizing the annotation of “demoralize” for its negative Propriety, we ensure that the primary evaluative stance of the sentence is captured. This approach aligns with our principle of marking the smallest unit that conveys the overall attitude, emphasizing the importance of clarity in expressing evaluative meanings. Not annotating *seeks to* helps to avoid potential conflicts in evaluative polarity (positive vs. negative) that could arise from annotating both expressions. This decision ensures that our annotations remain coherent and focused on the most salient evaluative aspects of the text.

A second challenge encountered during the annotation process concerned the decision on how many segments should be annotated within a single sentence. This challenge is exemplified by (3) below.

(3-1) ***No country has as much diplomatic clout with Russia** (+Capacity) **while also having equally good ties with Ukraine as China** (+Capacity).*

(3-2) ***No country has as much diplomatic clout with Russia while also having equally good ties with Ukraine as China** (+Capacity).*

The example presents a comparative assessment of China’s diplomatic ties with Russia and Ukraine,

leading to a question: Should this be annotated as exhibiting one instance of Capacity that covers the entire comparative structure, or as two separate instances of Capacity for each of the diplomatic relationships mentioned? We eventually decided on a single annotation for the unified concept approach. The decision to annotate sentence (3-2) as one instance stems from the recognition that the sentence articulates a singular, overarching evaluative stance regarding China’s diplomatic capabilities. The comparative structure of the sentence suggests a holistic evaluative judgement rather than two distinct evaluations. It reflects the integrated nature of the evaluative statement, where the two aspects of China’s diplomatic relations are not isolated evaluations but interconnected to produce a singular assessment.

A further challenge concerns the appropriate scope for annotating evaluative meanings, especially when a single term might embody the evaluation, but its full implication becomes apparent only in a broader context. This challenge is illustrated by example (4).

(4-1) *Over the past decade, Russia had gradually **transformed** (+Capacity) itself from a marginal player in Asian affairs into a potential “third force” amid rising Sino-US rivalry.*

(4-2) *Over the past decade, Russia had gradually **transformed itself from a marginal player in Asian affairs into a potential “third force” amid rising Sino-US rivalry** (+Capacity).*

Here, *transformed* implies a significant and beneficial change, potentially warranting an annotation as Capacity on its own. However, the broader context provided by the complete phrase offers a more comprehensive understanding of Russia’s change in status. To address this, we decided to annotate the entire phrase *transformed itself from a marginal player in Asian affairs into a potential “third force” amid rising Sino-US rivalry* as Capacity in (4-2). This decision is based on the understanding that the full evaluative impact of Russia’s transformation is most accurately captured when considering the entire phrase. This approach allows for a more precise capture of the evaluative meaning, acknowledging that the significance of the transformation encompasses not just the act of change (*transformed*) but its direction and outcome (from a marginal player to a potential “third force”).

#### 4.2 Challenges in Classifying Appraisal and Possible Solutions

We encountered significant challenges during the practical implementation of classification. These challenges primarily stem from the inherent subjectivity in distinguishing attitudes and the

vague boundaries between different evaluative categories. These factors frequently led to discrepancies among annotators, underscoring the need for a refined approach to ensure consistency and transparency in the annotation.

A major challenge is found in distinguishing between Affect and Appreciation, which is particularly pertinent when considering categories such as Security (a subcategory of Affect) versus Reaction (a subcategory of Appreciation). Consider

- (5-1) *Putin's position, and perhaps his life, is at **risk** (-Security) if there is another big Ukrainian victory.*
- (5-2) *Putin's position, and perhaps his life, is at **risk** (-Reaction) if there is another big Ukrainian victory.*

It could be argued that the phrase *at risk* should be classified under Affect, focusing on Security as it highlights concerns for Putin's personal safety and political stability. This interpretation emphasizes the emotional impact and the sense of threat to well-being, suggesting Affect as the fitting category. Alternatively, the same phrase could be analyzed as Appreciation with an emphasis on Reaction. This analysis assesses the sentence as evaluating the consequences or outcomes of a potential event on Putin's position, considering it an evaluation of situational change rather than an emotional response. The choice between Affect and Appreciation thus hinges on the interpretation of the sentence's core focus. If viewed primarily as eliciting an emotional response regarding Putin's precarious situation, Affect is deemed appropriate. However, if the sentence is interpreted as assessing the impact of potential events on Putin's status, Appreciation would be chosen.

The differentiation between Affect and Judgement presents another layer of complexity in the annotation process, especially when sentences can potentially align with either category based on their evaluative focus. This challenge is illuminated in (6).

- (6-1) *Ukrainians **have needlessly suffered a terrible toll** (-Happiness) and the impact is rippling around the world, with disruptions to food supplies and higher energy and grain costs bringing hunger and poverty to tens of millions of vulnerable people.*
- (6-2) *Ukrainians **have needlessly suffered a terrible toll** (-Propriety) and the impact is rippling around the world, with disruptions to food supplies and higher energy and grain costs bringing hunger and poverty to tens of millions of vulnerable people.*

Opting to annotate as "-Happiness" suggests an interpretation focused on the emotional response elicited by the Ukrainians' suffering, reflecting the emotional distress and negative states, hence fitting the Affect category. Alternatively, a perspective on Propriety shifts the perspective towards a moral or ethical Judgement. This view interprets it as a violation of moral standards, emphasizing the situation's ethical implications over its emotional impact.

The ambiguities between Affect vs. Judgement and Affect vs. Appreciation are a notable challenge that has been identified in the literature. Thompson (2014) has referred to this as the "Russian doll effect", where evaluative meanings are nested within one another, potentially qualifying for multiple categories of appraisal. Double annotation has been advocated to capture the layered nuances of evaluative language (Macken-Horarik and Isaac, 2014). However, for the sake of consistency and simplicity in the annotation process, we decided to adhere to a single annotation and to categorize based on the most prominent aspects: Affect for evaluations relating to emotions or feelings of people, Judgement for behaviours or actions, and Appreciation for objects or phenomena.

A single annotation approach simplifies the process, making it more accessible and manageable for annotators. Double annotation, while potentially offering a richer analysis, introduces complexity that could hinder the efficiency and consistency of the annotation process. The single annotation can be supplemented by a comprehensive textual analysis at a later stage, which will allow for a deeper exploration of the texts, where the nuances that might have been simplified during the annotation can be revisited and analyzed in greater depth. This strategy does not overlook the complexity of evaluative language but rather postpones a more granular analysis to the post-annotation stage. Here, the annotations serve as a foundation for further reflection and investigation, allowing researchers to explore the "Russian doll effect" with the full context of the text in view. This reflective analysis enables us to understand how evaluative meanings are interwoven and how they contribute to the overall discourse.

We have discussed the primary distinctions among Affect, Judgement and Appreciation. These distinctions are critical for identifying the broad categories in which language can express evaluations and attitudes. Finer distinctions need to be investigated, especially in Judgement. Consider

- (7-1) *The cold shoulder: Richard Heydarian says the Ukraine invasion **has soured Russia's ties across Southeast Asia** (-Normality).*

(7-2) *The cold shoulder: Richard Heydarian says the Ukraine invasion **has soured Russia's ties across Southeast Asia** (-Propriety).*

(7-1) is labelled as -Normality, suggesting that the Ukraine invasion is being evaluated in terms of its deviation from expected or conventional diplomatic behaviour, thus affecting Russia's international relationships. The focus is on the abnormality of the situation, implying that such actions are not in line with what is typically expected in international relations, leading to a deterioration in ties. Alternatively, it can also be annotated as negative Propriety in (7-2), shifting the emphasis to the appropriateness of the invasion and its consequences. This perspective assesses the invasion's impact on diplomatic relationships as a matter of ethical judgement, suggesting that the action is morally wrong or unacceptable, hence the negative repercussions on Russia's relations. Given the nuanced differences between Normality and Propriety within the Judgement category, where Normality is associated with social esteem and Propriety with social sanction, the challenge arises in ensuring accurate and consistent annotation.

To address this challenge and enhance both inter-rater agreement and consistency, we decided to prioritize Propriety when overlapping occurs. When an evaluative statement could potentially be annotated as both Propriety and Normality, the guidelines should advise annotators to prioritize Propriety. The prioritization is grounded in the intrinsic relationship and hierarchy between these concepts. Propriety encompasses appropriateness, which inherently requires actions or behaviour to align with societal norms and expectations, thus implying Normality. However, Normality focuses solely on the conformity of actions with norms and standards without necessarily engaging with their moral or ethical dimensions. Propriety assessments include a judgement of Normality but also extend beyond to consider legal appropriateness. By adopting Propriety as the default category in cases of overlap, annotators are likely to achieve higher consistency in their evaluations.

Distinguishing between Capacity and Tenacity within the Judgement category presents another layer of complexity. Both subcategories pertain to evaluations of behavior, but they focus on different aspects. Capacity refers to the ability or power to do something, often related to skill or competence. Tenacity, on the other hand, emphasizes persistence or determination in pursuing goals, especially in the face of obstacles.

(8-1) *Through its permanent seat on the United Nations Security Council, it also has the means to **ensure that countries adhere to global standards** (+Capacity).*

(8-2) *Through its permanent seat on the United Nations Security Council, it also has the means to **ensure that countries adhere to global standards** (+Tenacity).*

To navigate the distinction between Capacity and Tenacity more effectively in (8), we have to carefully examine the context to identify whether the emphasis is on the inherent ability (Capacity) or on the persistence and determination (Tenacity). Tenacity often implies a sustained effort in the face of challenges or obstacles. If the text highlights overcoming difficulties or persistent effort, Tenacity might be the more appropriate category. In contrast, Capacity focuses on the ability or competence without necessarily implying effort against resistance.

An additional issue is the disproportionate representation of the Valuation subcategory within Appreciation compared to Reaction and Composition. As illustrated in Figure 3, the instances of Valuation across the four newspapers significantly outnumber those of the other two subcategories within Appreciation. It emerged during the annotation that when segments did not clearly align with Reaction or Composition, there was a tendency to categorize them as Valuation.

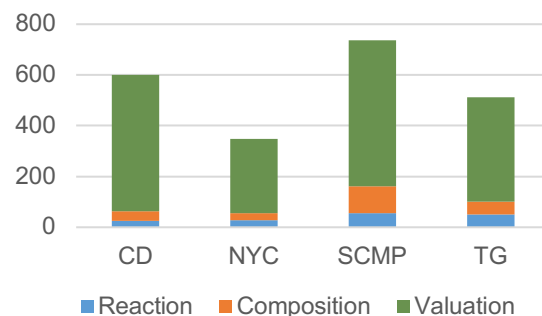


Figure 3: Distribution of subcategories under Appreciation for different newspapers

The use of Valuation as a catch-all category, while streamlining the annotation process, introduced challenges in analysis. The over-representation of Valuation could dilute the specificity of our findings, making it harder to discern distinct patterns or nuances in evaluative expressions. Such a broad categorization risks oversimplifying the rich evaluative landscape present in discourse, potentially masking the intricate ways in which objects or phenomena are appraised. To address this issue, it is crucial to refine the criteria for categorizing evaluative expressions under Valuation. This refinement process may involve expanding the Appreciation dimension to include additional, more specific categories tailored to the texts being analyzed. By doing so, we can accommodate a broader range of evaluative expressions, ensuring a more granular and accurate classification. Thus, while annotating, there is an opportunity to extend the Appreciation

categories as needed, ensuring that the framework remains flexible and responsive to the complexities of the texts under examination.

## 5. Conclusion

Throughout this project, we explored the practical application of Appraisal Theory in the task of corpus annotation with a particular focus on the Attitude system. The endeavour was driven by the aim to illuminate the complexities and challenges inherent in the annotation process and to come up with effective strategies for overcoming these obstacles. Central to our project was the formulation and implementation of a set of annotation guidelines to ensure accuracy and consistency. These principles guided our approach to identifying evaluative expressions, considering their contextual implications and categorizing them accordingly. Through this practical methodology, we aimed to refine the process of corpus annotation, making it a more effective tool for semantic annotation in general and pragmatic annotation of stance in particular.

Our annotation experiment revealed significant issues, particularly in the dual tasks of identifying and classifying evaluative expressions within the texts, highlighting the complexity of Appraisal Theory and the inherent subjectivity in interpreting expressions of Attitude. Our corpus-informed solutions involved a detailed examination of the Attitude category in authentic texts, leading to the formulation of strategies to resolve specific confusable annotations. This approach facilitated a more structured annotation process contributing to the broader issue of semantic and pragmatic annotation of corpus data involving subjective judgements. Moreover, this study identified a need for flexibility within the annotation framework, especially in addressing the disproportionate use of the valuation subcategory within Appreciation. This observation prompted a critical re-evaluation of our classification strategy, allowing for a more granular analysis of evaluative language.

During the refinement of our annotation guidelines within the Appraisal Theory framework, we realized that incorporating parts of speech (POS) and phrasal structures into our definitions of annotation units had not been explicitly stated. Addressing this could substantially enhance the degree of transparency and consistency in the identification of evaluative segments. Insights from Caro (2014) and Hunston and Su (2019), who emphasize the evaluative potential of adjectives, nouns and verbs, suggest that future efforts could adopt a hierarchical approach to annotation. Such an approach would give precedence to adjective phrases due to their prominent role in conveying evaluative meaning while still recognizing the contributions of nouns and verbs. Additionally, it was decided that nouns derived from adjectives and verbs should be

annotated accordingly. Our updated strategy for selecting the smallest text segment for annotation advocates a flexible method: starting with single lexical items, then expanding to phrases, and eventually to clauses if necessary. Future enhancements to our annotation guidelines might also benefit from including phrasal and syntactic structures. Acknowledging the syntactic roles of adjectives, nouns, and verbs within their respective phrases could help more precisely to identify the scope of evaluative expressions. It should be noted that while the discussions so far have centred on grammatical aspect, semantic factors are fundamentally important and form a major basis of annotation judgements.

In conclusion, our experiment reported and addressed the practical task of annotating pragmatic information within the framework of Attitude in Appraisal Theory. It has detailed the process of identifying and classifying evaluative expressions, thereby enhancing both transparency and consistency in the annotation practices. By proposing specific solutions to the intricacies involved in the annotation of evaluative language, this work contributed towards the methodological foundations for future research. Our future work will incorporate parts of speech and phrasal structures into annotation guidelines and adopt a hierarchical approach to better capture evaluative text segments. We plan to integrate parts of speech and phrasal structures into our annotation guidelines, employing a hierarchical approach to identify evaluative text segments more consistently in conjunction with semantic considerations.

## 6. Acknowledgement

This work was supported in part by grants received from China's National Planning Office of Philosophy and Social Sciences (Project No 22BYY009), Beijing Social Sciences Foundation (Project No. 18JDYYA005) and City University of Hong Kong (Project Grant Nos 7020036, 9360115 and 6008167). The second author would like to thank the Halliday Centre for Intelligent Applications of Language Studies at City University of Hong Kong for a visiting professorship received in 2023.

## 7. Bibliographical References

- Bednarek, M. (2006). *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. Continuum.
- Bednarek, M. (2009). Language patterns and attitude. *Functions of Language*, 16(2), 165–192.
- Bednarek, M., & Caple, H. (2010). Playing with environmental stories in the news – Good or bad practice? *Discourse & Communication*, 4, 5–31.

- Breit, B. W. (2014). Appraisal theory applied to the wine tasting sheet in English and Spanish. *Ibérica, Revista de la Asociación Europea de Lenguas para Fines Específicos*, (27), 97-120.
- Caro, E. M. (2014). The expression of evaluation in weekly news magazines in English. In G. Thompson & L. Alba-Juez (Eds.), *Evaluation in context* (pp. 321–343). Amsterdam and Philadelphia: John Benjamins.
- Fuoli, M. (2018). A stepwise method for annotating APPRAISAL. *Functions of Language*, 25(2), 229-258.
- Fuoli, M., & Hommerberg, C. (2015). Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions. *Corpora*, 10(3), 315-349.
- Geng, Y., & Wharton, S. (2016). Evaluative language in discussion sections of doctoral theses: Similarities and differences between L1 Chinese and L1 English writers. *Journal of English for Academic Purposes*, 22, 80-91.
- Hood, S. (2010). *Appraising research: Evaluation in academic writing*. Springer.
- Huan, C. (2016). Journalistic engagement patterns and power relations: Corpus evidence from Chinese and Australian hard news reporting. *Discourse & Communication*, 10(2), 137-156.
- Hunston, S., & Su, H. (2019). Patterns, constructions, and local grammar: A case study of 'evaluation'. *Applied Linguistics*, 40(4), 567-593.
- Macken-Horarik, M., & Isaac, A. (2014). Appraising appraisal. In G. Thompson & L. Alba-Juez (Eds.), *Evaluation in context* (pp. 67–92). Amsterdam and Philadelphia: John Benjamins.
- Martin, J. R. (2000). Beyond exchange: Appraisal systems in English. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 142–175). Oxford: Oxford University Press.
- Martin, J. R. and White, P. R. R. (2005). *The Language of Evaluation: Appraisal in English*. London: Palgrave.
- Mayo, M. A., & Taboada, M. (2017). Evaluation in political discourse addressed to women: Appraisal analysis of Cosmopolitan's online coverage of the 2014 US midterm elections. *Discourse, context & media*, 18, 40-48.
- Meadows, B., & Sayer, P. (2013). The Mexican sports car controversy: An appraisal analysis of BBC's Top Gear and the reproduction of nationalism and racism through humor. *Discourse, Context & Media*, 2(2), 103-110.
- O'Donnell, M. (2008). Demonstration of the UAM CorpusTool for text and image annotation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pp. 13–16. Association for Computational Linguistics.
- Pounds, G. (2011). This property offers much character and charm: Evaluation in the discourse of online property advertising. *Text & Talk*, 31(2), 195–220.
- Swain, E. (2010). *Getting engaged: Dialogistic positioning in novice academic discussion writing*. EUT Edizioni Università di Trieste.
- Taboada, M., Carretero, M., & Hinnell, J. (2014). Loving and hating the movies in English, German and Spanish. *Languages in Contrast*, 14(1), 127-161.
- Thompson, G. (2014). Affect and emotion, target-value mismatches, and Russian dolls: Refining the appraisal model. In G. Thompson & L. Alba-Juez (Eds.), *Evaluation in context* (pp. 47–66). Amsterdam and Philadelphia: John Benjamins.



# Attractive Multimodal Instructions

## Describing Easy and Engaging Recipe Blogs

Ielka van der Sluis, Jarred Kiewiet de Jonge  
Center for Language and Cognition Groningen (CLCG)  
University of Groningen The Netherlands  
{i.f.van.der.sluis, j.j.b.kiewiet.de.jonge}@rug.nl

### Abstract

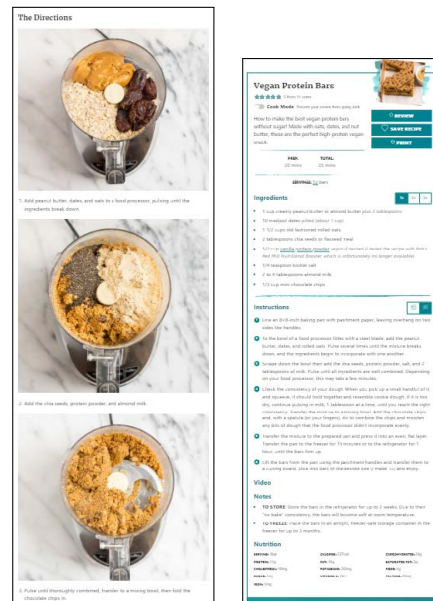
This paper presents a corpus study that extends and generalises an existing annotation model which integrates functional content descriptions delivered via text, pictures and interactive components. The model is used to describe a new corpus with 20 online vegan recipe blogs in terms of their Attractiveness for at least two types of readers: vegan readers and readers interested in a vegan lifestyle. Arguably, these readers value a blog that shows that the target dish is Easy to Make which can be inferred from the number of ingredients, procedural steps and visualised actions, according to an Easy to Read cooking instruction that displays a coherent use of verbal and visual modalities presenting processes and results of the cooking actions involved. Moreover, added value may be attributed to invitations to Engage with the blog content and functionality through which information about the recipe, the author, diet and nutrition can be accessed. Thus, the corpus study merges generalisable annotations of verbal, visual and interaction phenomena to capture the Attractiveness of online vegan recipe blogs to inform reader and user studies and ultimately offer guidelines for authoring effective online multimodal instructions.

**Keywords:** multimodal instruction, document design, corpus analysis, vegan recipe blogs

## 1. Introduction

### 1.1. Multimodal Recipe Blogs

Recipes have been a source of inspiration for structured text analysis for some time now (Bieñ et al., 2020; DiMeo and Pennell, 2018; Floyd and Forster, 2017; Mori et al., 2012; Görlach, 1992). In addition, recipes are often composed of verbal and visual modes and thus allow for the evaluation of the effectiveness of multimodal presentations for a variety of readers as well as users in multiple respects (e.g., attractiveness, comprehension, performance). Recipe blogs are a specific type of online documents that share recipes, cooking tips and food-related content. A recipe blog presents a procedural instruction that guides users through the steps involved to prepare a dish (Van der Sluis and Mellema, Submitted). Figure 1 illustrates that recipe blogs present the instruction in two formats (Bowker, 2021; Domingo et al., 2014). The blog as a whole presents an Instruction with Pictures (IWP), a



(a) IWP of MI 3.

(b) RC of MI 3.

Figure 1: Source: <https://www.wellplated.com/vegan-protein-bars/>.

multimodal step-by-step instruction combining text and pictures, and allows for additional dynamic content (e.g., adds, videos). At the end of the blog a Recipe Card (RC) is offered, which presents all the necessary steps and ingredients to prepare the dish in text.

With the growing popularity of mindful dieting, a large body of blogs offer recipes for crafting nutritious and healthy dishes at home (Guha and Gupta, 2020), but what makes a blog attractive? The study presented in this paper examines the means that authors of online content use to attract their public. Retrieval and analysis of blogs are interesting because authors simultaneously employ a range of semiotic elements (e.g., text, pictures, videos, interactive features). To explore the blog authors' use of available modes and functions to attract potential online readers and users, we conducted a small and focused corpus study. Based on existing approaches and findings in multimodal and online content analysis, the corpus study is offered as a starting point to conduct future reader and user evaluations and to support the further development and automation of our preliminary notion of Attractiveness.

The corpus solely contains recipes for vegan nutrition bars ie. compact and portable snacks typically crafted from plant-based ingredients like nuts, seeds, fruits, and grains that serve as a source of essential nutrients catering to health-conscious and environmentally-aware consumers. Studying vegan blogs is timely because the past decade displays a noticeable shift and steady increase in the adoption of an exclusively plant-based lifestyle (Kustar and Patino-Echeverri, 2021; Kamiński et al., 2020; Schösler et al., 2012). In the cooking domain, this trend is mirrored in recipe blogs that support a vegan diet (Asano and Biermann, 2019). Given the abundance of recipe blogs and the variation in which the food preparation procedures in them are presented it is of interest to identify the characteristics that make a blog attractive to both vegans as well as those that are merely interested in a vegan lifestyle. Recipes for vegan nutrition bars in particular offer potential to convince blog users due to their convenience and popularity as a healthy and

nutritious snack or meal replacement option that can contribute to supporting a healthier diet in a relatively quick and easy way (Bansal et al., 2022; Jovanov et al., 2021).

The corpus study was set up to answer the following research question: Which means do authors of step-by-step recipe blogs for vegan nutrition bars use to attract potential users? Sections 1.2 and 1.3 introduce the background for a notion of Attractiveness which is operationalised using three aspects: Easy to Make, Easy to Read and Engagement.

## 1.2. Attractive Vegan Recipe Blogs

Arguably, recipes are Easy to Make dependent on the number of ingredients involved, the number of procedural steps described and the availability of visual presentations of those steps. The recipe becomes Easy to Read when the instructional parts of the blog display coherence and consistency in terms of their text content (RC vs. IWP) and coherence in the text-picture combinations within the IWP (cf. Bowker, 2021; Li and Xie 2020; Kang, 2010). Engagement requires alignment of the author's values with the needs and preferences of the blog users (Cooper et al., 2022; Machnee, 2019) as well as useful and playful content (Mainolfi et al., 2022; Liao et al., 2013). Given the wealth of online food recipes, blog authors are compelled to grab the attention of blog users. Independent of the intrinsic qualities of a recipe, the presence and professionalism of the blog pictures is crucial in influencing readers to choose a recipe (Starke et al., 2021), although effects of visual content, colourfulness, appearance of human faces and text-picture relations depend on the social medium platform (Li and Xie, 2020). Apart from quality pictures, the presence and credibility of the blog author is important. Bloggers should be knowledgeable, influential, passionate, transparent and reliable (Kang, 2010; Rubin and Liddy, 2006). A blogger's appearance hinges on a learned, positive writing style while credibility, trust and authenticity are gained through sharing personal stories (Machnee, 2019). At last, users increasingly consider nutritional characteristics when

selecting recipes to support informed decisions that align with their health and dietary needs (Cooper et al., 2022; Cheng et al., 2021; Rokicki et al., 2018; Trattner et al., 2018; Elswailer et al., 2017; Van Pinxteren et al., 2011; Freyne and Berkovsky, 2010). Accordingly, in a recipe blog the information about nutrition and diet should be present and easy to find.

### 1.3. Annotating Multimodal Instructions

Multimodality requires interdisciplinary research because multiple modes cohere and make meaning together (Bateman et al., 2017; Jewitt, 2009). Multimodal recipes rely on a combination of textual directions and visual cues (Ganier, 2012, 2000; Mayer, 2005). An instructive text assists people in executing a task through a step-by-step description of procedural information, usually presented in a numbered list of actions (Karreman and Loorbach, 2013). Alongside the procedural information instructions also contain control information (Van der Sluis et al., 2022; Karreman et al., 2005), encompasses non-procedural supplementary details relevant to the described process such as warnings, explanations, conditions etc. In instructions these two types of information work in tandem to ensure that users have the necessary knowledge and understanding to complete a task successfully (Ummelen, 1997).

Bateman (2014) describes coherence relations between text and pictures in terms of how one mode expands the meaning of the other (cf. Van der Sluis and Mellema, Submitted; Halliday and Matthiessen, 2013; Kress and Van Leeuwen, 2001; Barthes, 1977). Elaboration occurs when information is restated in another mode at a similar level of generality. For instance, an action is described in the text as a process (e.g., mix ingredients) and the related picture presents the result of that action (e.g., the dough as a result from mixing the ingredients). Enhancement on the other hand involves providing qualifying information related to aspects such as time, place, manner, reason, purpose, and other circumstantial

restrictions. For instance, the text describes an action (e.g., stir a substance) and the picture shows that the action is performed using a particular utensil (e.g., a whisk is used to stir a substance).

In multimodal instructions procedures can be described in terms of the actions involved. Recently, human action annotation and retrieval gained interest in multiple domains, media and applications (Pustejovsky and Krishnaswamy, 2022; Alikhani et al., 2019; Pustejovsky, 2018; Van der Sluis et al., 2018; Pustejovsky et al., 2017; Zhang et al., 2016; Lev et al., 2016; Laptev et al., 2008). The annotation model proposed to describe the vegan nutrition bar recipe blog corpus employs and extends the action-based PAT annotation model (Van der Sluis et al., 2022, 2017, 2016b)<sup>1</sup>. The PAT model has been used to describe (parts of) multimodal instructions according to the following steps:

1. The instructional text is split into clauses;
2. The clauses are identified as either Action clauses or Control Information clauses;
3. The text clauses and the accompanying instructional pictures are described using functional attributes (e.g., Action Type, Action Status, Action Aspect, Control Information, Specification);
4. Coherence relations are described as compositions of text and picture annotations.

The generalisability of the PAT model is shown by annotating multimodal instructions in different domains, such as first-aid instructions (Van der Sluis et al., 2017) and cooking instructions (Van der Sluis and Mellema, Submitted; Van der Sluis et al., 2016b), through the annotation of multiple document types e.g., illustrated texts; instructional videos (Vijfvinkel et al., 2018) and instructional comics (Wildfeuer et al., 2022). The current corpus study presents a further development which merges

---

<sup>1</sup>In the Pictures And Text or PAT project (<https://www.rug.nl/let/pat>), the PAT workbench (Van der Sluis and Redeker, 2019; Van der Sluis et al., 2016a) was built as an online tool designed to systematically describe multimodal documents.

annotation of different phenomena i.e. text, pictures, text-picture relations and interaction components to achieve a description of a context dependent notion of Attractiveness while further exploring the model's generalisability by describing online multimodal instructions.

## 2. Method

### 2.1. Corpus

The online recipe blogs for vegan nutrition bars were collected according to the following selection criteria:

- the blog includes an IWP and a RC;
- the IWP text describes the cooking procedure in at least three steps;
- the IWP includes at least three pictures visualising different stages in the cooking procedure;
- the RC has at least three procedural steps.

The vegan nutrition bar corpus contains 20 online recipes that were derived from eight distinct sources to allow a comparison between recipes from the same website while also ensuring a diverse representation across multiple sources. The corpus consists of four distinct parts with 5 recipes each: Part 1 contains 5 blogs from 5 different websites: Eat with Clarity<sup>2</sup>, Vegan Huggs<sup>3</sup>, Well Plated<sup>4</sup>, Hummusapien<sup>5</sup>, and Minimalist Baker<sup>6</sup>. Part 2, 3 and 4 contain 5 recipes respectively from Veggie World<sup>7</sup>, All-Purpose Veggies<sup>8</sup>, Eating Bird Food!<sup>9</sup>.

### 2.2. Annotation Model

The corpus study was set up to answer the following research question: Which means do authors of step-by-step recipe blogs for vegan

nutrition bars use to attract potential users? Attractiveness is operationalised using three aspects: Easy to Make, Easy to Read and Engagement, where the description of the notions Easy to Make and Easy to Read applies to particular parts of the blog namely the Instruction with Pictures and the Recipe Card, while the description of Engagement applies to the blog as a whole. The annotation model was largely based on the findings discussed in Section 1 of this paper. The annotation model was crafted and applied by two annotators that improved their work through multiple rounds of discussions until they agreed on the resulting model and the corpus description.

#### 2.2.1. Easy to Make and Easy to Read

Conceivably, food preparation becomes or appears easier when a recipe includes only a few ingredients, when the procedure includes only a few steps and when the steps are visualised (cf. Yajima and Kobayashi, 2009). The blogs are described accordingly, using a notion Easy to Make that includes: (1) the number of necessary ingredients; (2) the number of steps in which the procedure is presented in the IWP; and (3) the number of visualisations of the procedural steps presented in the text.

A recipe becomes Easier to Read when the presentation in the instructional parts of the recipe blog displays coherence in terms their text content as well as coherence in combining text and pictorial information (cf. Kang, 2010). An action-based approach was taken to describe the coherence of the Instruction with Pictures and the Recipe Card for each blog in the corpus. The models described by (Van der Sluis and Mellema, Submitted; Van der Sluis et al., 2016b) were used as a starting point. Table 1 presents the text, pictures, and text-picture relation categories. The text clauses are annotated as Action or Control Information (Van der Sluis et al., 2022) Action clauses and visualised actions in pictures are annotated in terms of Status (i.e. Obligatory, Alternative, Conditional) and Aspect (i.e. Process, Result), where the Aspect value in the pictures is dependent on whether any utensils are included

---

<sup>2</sup><https://eatwithclarity.com/>

<sup>3</sup><https://veganhuggs.com/>

<sup>4</sup><https://www.wellplated.com/>

<sup>5</sup><https://www.hummusapien.com/>

<sup>6</sup><https://minimalistbaker.com/>

<sup>7</sup><https://veggieworldrecipes.com/>

<sup>8</sup><https://allpurposeveggies.com/>

<sup>9</sup><https://www.eatingbirdfood.com/>

in the visualisation. The Control Information clauses include Warning, Condition, Manner, Advice, Explanation, Motivation, Purpose and Situation Sketch. The text-picture relations are described in terms of Layout (i.e. Index, Proximity) and Content (i.e. Enhancement, Elaboration). The content relations are described in terms of meaning expansions, given a particular action that is presented in the two modes (Bateman, 2014).

### 2.2.2. Engagement

Engagement is described in terms of the presence of the following Text, Picture and Interaction attributes in the blog as a whole.

The following Text attributes are described:

- Attention Grabber - introduction text that reels in the audience such as “These vegan protein bars are a cookie dough flavored treat you’re going to love” (MI R1).
- Author Welcome - explicit greeting from the authors e.g., “Hey there! We’re jasmine and chris” (MI R2).
- Diet Legend - keys that specify the diets for which the recipe is suitable (e.g., VG, V, DF for respectively Vegan, Vegetarian, Gluten free).
- Location Diet Legend - place in the blog where the Diet Legend is offered (Top, Bottom, NA).
- Nutrition Facts - alimentary types and quantities included in the recipe (i.e. fat, carbs, sugars, protein, vitamins and minerals).
- Location Nutrition Facts - place in the blog where the Nutrition Facts are offered (i.e. Top, Bottom).

Pictures are described as follows:

- Author Portrait - picture of the blogger.
- Teaser - picture of the end result.
- Ingredients - picture of prepped but uncooked ingredients.
- Recommendation - picture of other recipes.

Included Interaction aspects are:

- Jump to Recipe - button to go to the RC.
- Link to Author - pointer to blogger details.
- Social Handles - pointers to the blogger’s social media.

- Rate Option - evaluate the recipe on a scale.
- Comment Option - write recipe evaluation.
- Tick-off function - boxes to indicate that ingredients are handy.

## 3. Analysis

### 3.1. Easy to Make and Easy to Read

Table 2 presents an overall description of the four parts of the Vegan Nutrition Bar Corpus that indicate in how far the recipes are Easy to Make. The 5 recipes from Veggie World contain the most ingredients, steps and pictures compared to the other subsets in the corpus. The average number of steps and pictures are balanced within each of the corpus parts. The number of ingredients and the number of steps seem unrelated e.g., Part 1 and Part 4 include more ingredients than steps.

In terms of Easy to Read, Table 3 presents the frequencies and percentages of Action Status and Control Information in the IWPs and RCs. The corpus has 1020 clauses: 654 Action and 375 Control Information clauses. The RCs contain more clauses ( $N = 580$ ) than the IWPs ( $N = 440$ ), with similar distributions of Actions and Control Information within the IWPs and RCs (IWP  $\approx 63\%$  versus RC  $\approx 37\%$ ). Most Action clauses (IWP = 221; RC = 299) present Obligatory Actions. The most frequent Control Information clauses present the Manner in which to perform an action ( $N = 78$ ) and the Purpose for carrying out an action ( $N = 76$ ).

Table 4 presents the frequencies and percentages of Action and Control Information clauses in the four corpus parts. The number of Actions varies between the subsets with a maximum of 202 actions in Veggie World and a minimum of 105 in All-Purpose Veggies!. The sets do not vary much in the number of Control Information clauses ( $N \approx 94$ ).

Table 5 presents the text-picture relations in the IWPs in terms of Action Status and Action Aspect per corpus part. The IWPs contain 147 visualised actions and 279 verbalised actions. All pictures present Obligatory Actions, while the texts also contain Alternative ( $N = 27$ ) and Conditional Actions ( $N = 31$ ). The actions in





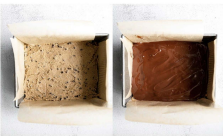


| Text Attribute      | Value       | Description                                                                                                                                                                     | Example (Source)                                                                                                                                       |
|---------------------|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| Action Status       | Obligatory  | An action that must be executed to perform the task successfully.                                                                                                               | "Melt the dark chocolate chips in a tall glass." (MI 10)                                                                                               |
|                     | Alternative | An action that can be executed as a replacement of another action.                                                                                                              | "(add more milk)...or water" (MI 14)                                                                                                                   |
|                     | Conditional | An action that can or must be executed under particular circumstances.                                                                                                          | "then coat in melted chocolate." (MI 11)                                                                                                               |
| Action Aspect       | Process     | The action is described as a process/in progress.                                                                                                                               | "Sprinkle with some flaky salt" (MI 18)                                                                                                                |
| Control Information | Warning     | The presentation addresses a possible danger.                                                                                                                                   | "Be careful to avoid burning the coconut" (MI 20)                                                                                                      |
|                     | Condition   | The presentation specifies a condition or circumstance for an action to be performed.                                                                                           | "Once your coconut has cooled," (MI 20)                                                                                                                |
|                     | Manner      | The presentation addresses the way in which an action must be executed.                                                                                                         | "until everything is evenly coated" (MI 19)                                                                                                            |
|                     | Advice      | The content of the presentation gives a recommendation on how to execute the action (not mandatory).                                                                            | "I suggest storing these vegan protein bars in the fridge" (MI 18)                                                                                     |
|                     | Explanation | The presentation offers more information on how to execute the action.                                                                                                          | "Each will give it a slightly different hue of green." (MI 6)                                                                                          |
|                     | Motivation  | The presentation addresses a positive feeling or action.                                                                                                                        | "and enjoy!" (MI 6)                                                                                                                                    |
| Picture Attribute   | Value       | Description                                                                                                                                                                     | Example (Source)                                                                                                                                       |
|                     | Obligatory  | An action that must be executed to perform the task successfully.                                                                                                               |  (MI R9)                                                            |
|                     | Process     | The action is visualised with utensils and/or human hands.                                                                                                                      |  (MI R5)                                                            |
|                     | Result      | The situation after completing an action, shown without utensils or hands.                                                                                                      |  (MI R9)                                                            |
| Relation Attribute  | Value       | Description                                                                                                                                                                     | Example (Source)                                                                                                                                       |
| Layout              | Index       | Picture and text are related via the use of numbers, letters or titles.                                                                                                         |  "4. Now add the mixture to the dates..." (MI R2)                  |
|                     | Proximity   | Picture and text are related because they are positioned near to each other and integrated in the text. Reading direction is more important than physical distance on the page. |  (MI R1)                                                          |
| Content             | Enhancement | Shows tools/hands to illustrate how the textualized action is performed.                                                                                                        |  "Whisk together the oat flour, protein powder and salt." (MI R1) |
|                     | Elaboration | Provides additional information, in terms of, provisions of a result state in specific details without tools/hands present.                                                     |  "Add on top of the bars"(MI R1)                                  |

Table 1: Easy to Read attributes to describe Text, Pictures and Text-Picture relations.

the IWP text are always verbalised as a Process. Visualised actions appear as a Process showing utensils (N = 86) or as a Result (N = 61), showing the derived end state of an action. The differences between the corpus parts are substantial; Veggie World employs mostly

Process visualisations, while the other subsets display more variation in Action Aspect.

Table 6 presents the Layout and Content relations between the IWP text and pictures. Indices are not used much in the corpus, only the Vegan Huggs recipe in Part 1 includes enumer-

| P   | Source | Ingredients | Steps | Pics |
|-----|--------|-------------|-------|------|
| 1   | 5      | 8.2         | 5.4   | 4.2  |
| 2   | 1      | 11.2        | 11.4  | 11.6 |
| 3   | 1      | 5.4         | 6.4   | 6.0  |
| 4   | 1      | 7.6         | 6.2   | 6.6  |
| All | 1      | 8.7         | 7.9   | 6.8  |

Table 2: Easy to Make - Number of Sources and averages for Ingredients, instructional Steps and instructional Pictures per corpus Part and in the whole corpus.

ation to relate the text and pictures. Elaboration relations between text and pictures ( $N = 90$ ), where the pictures present the result of a particular action are most frequent. Enhancement relations appear in 55 cases and mostly in the Veggy World blogs ( $N = 29$ ). Two pictures are not related to a clause, 15 clauses are related to more than one picture ( $N = 35$ ) and 899 clauses have no relation to any picture.

### 3.2. Engagement

Table 7 presents the frequencies and percentages for the Text, Picture and Interaction attributes to describe how authors invite user Engagement. In Text all blogs include Attention Grabbers and Nutrition Facts. In 2 of 20 blogs authors do not include an explicit welcome greeting. In 9 of 20 blogs the Diet Legend is omitted, which means that there is no indication about the suitability of the recipe for consumers with particular dietary constraints. The 9 Diet Legends that are included are always offered at the Top of the blog, while the Nutrition Facts are always offered at the bottom of the blog close to or as part of the RC.

The Picture attributes display that all blogs show a Teaser at the top of the blog that exemplifies the envisioned vegan nutrition bars. All blogs except one include a picture of the blog author. All blogs except one include Recommendations to other vegan recipes. Only 9 of 20 blogs include an image of the ingredients necessary to prepare vegan nutrition bars.

The Interaction attributes display that links to the Recipe card at the Bottom of the blog are included in all blogs. Also links to more

information about the author and the social media pages of the author are usually present. The means for a scaled or written evaluation of the recipes are also always included. The Tick-off function appears only in 9 of 20 blogs.

## 4. Discussion and Conclusion

The corpus study outlined in this paper offers a starting point to integrate different phenomena in online content with which the Attractiveness of online multimodal instructions that employ multiple modes (i.e. text, pictures, interaction components) can be described. The description provides a context dependent view on 20 vegan nutrition bar recipes from 8 sources constituted in three notions that offer insight in whether the blogs are Easy to Make, Easy to Read and Engagingly presented. In this case study the Attractiveness aspects were operationalised on the basis of existing findings from studies on multimodal communication, online content and the food domain, the newly developed categories may be complemented and improved in future work. For instance, in terms of Easy to make aspects such as preparation and cooking times or the availability of ingredients are likely of importance. In terms of Engagement, currently not all the attributes have equivalents in the described modalities. For example, the list of ingredients that is usually offered in the RC text was not included in the annotation model, while a picture of the ingredients was. Similarly, the nutritional facts are solely described as Text, while conceivably nutritional facts may also be visualised (cf. packaging of food products). Further grounding of the categories in terms of cultural and societal preferences are in order. For instance, the effectiveness of Process and/or Result visualisations in combination with verbalised Process actions needs further evaluation in a context of use, perhaps differentiating between novice and expert cooks. In the blog domain evaluation of the merit and/or annoyance of adds, videos and other dynamic content and of functions such as ticking off ingredients seems valuable. Finally, an extended description of author presence and credibility could be informed by

| Attribute        | Value       | IWP  |       | RC   |       | Total |       |
|------------------|-------------|------|-------|------|-------|-------|-------|
|                  |             | N    | %     | N    | %     | N     | %     |
| Action Clauses   | Obligatory  | 221  | 79.2% | 299  | 51.6% | 520   | 51.0% |
|                  | Alternative | 27   | 9.7%  | 33   | 5.7%  | 60    | 5.9%  |
|                  | Conditional | 31   | 11.1% | 34   | 5.9%  | 65    | 6.4%  |
| Total            |             | 279  | 63.4% | 366  | 63.1% | 654   | 63.2% |
| CI Clauses       | Manner      | 29   | 6.6%  | 49   | 8.4%  | 78    | 7.6%  |
|                  | Purpose     | 35   | 8.0%  | 41   | 7.6%  | 76    | 7.5%  |
|                  | Condition   | 25   | 5.7%  | 32   | 5.5%  | 57    | 5.6%  |
|                  | Advice      | 24   | 5.5%  | 33   | 5.7%  | 57    | 5.6%  |
|                  | Warning     | 18   | 4.1%  | 20   | 3.4%  | 38    | 3.7%  |
|                  | Motivation  | 15   | 3.4%  | 17   | 2.9%  | 32    | 3.1%  |
|                  | Explanation | 12   | 2.7%  | 17   | 2.9%  | 29    | 2.8%  |
| Situation Sketch | 3           | 0.7% | 5     | 0.9% | 8     | 0.8%  |       |
| CI Total         |             | 161  | 36.6% | 214  | 36.9% | 375   | 36.8% |
| Clause Total     |             | 440  | 100%  | 580  | 100%  | 1020  | 100%  |

Table 3: Easy to Read - Frequencies and percentages of Action and Control Information clauses in IWPs and RCs.

|                | Part | IWP |       | RC  |       | Total |       |
|----------------|------|-----|-------|-----|-------|-------|-------|
|                |      | N   | %     | N   | %     | N     | %     |
| Action Clauses | 1    | 46  | 10.5% | 117 | 20.2% | 163   | 16.0% |
|                | 2    | 91  | 20.7% | 111 | 19.1% | 202   | 19.8% |
|                | 3    | 52  | 11.8% | 53  | 9.1%  | 105   | 10.3% |
|                | 4    | 90  | 20.5% | 85  | 14.7% | 175   | 17.2% |
|                | All  | 279 | 63.4% | 366 | 63.1% | 645   | 63.2% |
| CI Clauses     | 1    | 30  | 6.8%  | 72  | 12.4% | 102   | 10.0% |
|                | 2    | 39  | 8.9%  | 56  | 9.7%  | 95    | 9.3%  |
|                | 3    | 46  | 10.5% | 39  | 6.7%  | 85    | 8.3%  |
|                | 4    | 46  | 10.5% | 47  | 8.1%  | 93    | 9.1%  |
|                | All  | 161 | 36.6% | 214 | 36.9% | 375   | 36.8% |
|                |      | 440 | 100%  | 580 | 100%  | 1020  | 100%  |

Table 4: Easy to Read - Frequencies and percentages of Action and Control Information clauses in IWPs and RCs per corpus Part.

profile factors like expertise, identity disclosure, trustworthiness, content quality and personal appeals (Rubin and Liddy, 2006).

Although the two annotators that conducted the study used various rounds in which the annotation model and the corpus description was discussed and improved, the effort needs further evaluation in terms of inter-annotator agreement to obtain an indication of the difficulty of the annotation task and to examine in which ways the model can be improved, complemented and made generalisable as to apply to other online instructive blog content. In addition, prompting large language models on

the classification and generation of attractive instructions could further strengthen the exploratory results offered in this paper. Large databases containing cooking instructions, as well as videos of people executing them (e.g., Yagcioglu et al., 2018; Carvalho et al., 2018; Salvador et al., 2017; Regneri et al., 2013; Rohrbach et al., 2012a; Rohrbach et al., 2012b) demonstrate that a combination of text-based models with visual information can significantly improve the understanding and assessment of action descriptions. Recent initiatives in natural language processing and generation are promising (e.g., Tu et al., 2022; Pustejovsky



| Attribute     | Value       | Part | IWP Text |       | IWP Pictures |       |
|---------------|-------------|------|----------|-------|--------------|-------|
|               |             |      | N        | %     | N            | %     |
| Action Status | Obligatory  |      | 221      | 79.2% | 147          | 100%  |
|               | Alternative |      | 27       | 9.7%  | 0            | 0%    |
|               | Conditional |      | 31       | 11.1% | 0            | 0%    |
| AS Total      |             |      | 279      | 100%  | 147          | 100%  |
| Action Aspect | Process     | 1    | 46       | 100%  | 12           | 46.2% |
|               |             | 2    | 91       | 100%  | 55           | 94.8% |
|               |             | 3    | 52       | 100%  | 6            | 20.0% |
|               |             | 4    | 90       | 100%  | 13           | 39.4% |
|               |             | All  | 279      | 100%  | 86           | 58.5% |
|               | Result      | 1    | 0        | 0.0%  | 14           | 53.8% |
|               |             | 2    | 0        | 0.0%  | 3            | 5.2%  |
|               |             | 3    | 0        | 0.0%  | 24           | 80.0% |
|               |             | 4    | 0        | 0.0%  | 20           | 60.6% |
|               |             | All  | 0        | 0.0%  | 61           | 41.5% |
| AA Total      |             |      | 279      | 100%  | 147          | 100%  |

Table 5: Easy to Read - Frequencies and percentages of Action Status and Aspect in IWP Text and Pictures.

| Part | Layout |      |           |       | Content     |       |             |       |
|------|--------|------|-----------|-------|-------------|-------|-------------|-------|
|      | Index  |      | Proximity |       | Enhancement |       | Elaboration |       |
|      | N      | %    | N         | %     | N           | %     | N           | %     |
| 1    | 7      | 4.8% | 19        | 13.1% | 11          | 7.6%  | 15          | 10.3% |
| 2    | 0      | 0%   | 58        | 40.0% | 29          | 20.0% | 29          | 20.0% |
| 3    | 0      | 0%   | 28        | 19.3% | 6           | 4.1%  | 22          | 15.2% |
| 4    | 0      | 0%   | 33        | 22.8% | 9           | 6.2%  | 24          | 16.6% |
| All  | 7      | 4.8% | 138       | 95.2% | 55          | 37.9% | 90          | 62.1% |

Table 6: Easy to Read - Frequencies and percentages of Layout and Content relations per corpus part.

| Category    | Attribute         | N  | %    |
|-------------|-------------------|----|------|
| Text        | Attention Grabber | 20 | 100% |
|             | Author Welcome    | 18 | 90%  |
|             | Diet Legend       | 11 | 55%  |
|             | Nutrition Facts   | 20 | 100% |
| Picture     | Author Portrait   | 19 | 95%  |
|             | Teaser            | 20 | 100% |
|             | Ingredients Pic   | 9  | 45%  |
|             | Recommendation    | 19 | 95%  |
| Interaction | Jump to Recipe    | 20 | 100% |
|             | Link to Author    | 19 | 95%  |
|             | Social Handles    | 20 | 100% |
|             | Rate Option       | 20 | 100% |
|             | Comment Option    | 20 | 100% |
|             | Tick-off Function | 9  | 45%  |

Table 7: Engagement - Frequencies and percentages for Text, Picture and Interaction attributes for the whole corpus.

et al., 2021). Thus, annotation of cooking instructions serves to build systems that understand and extract practical knowledge from written instructions, enabling them to offer guidance or to perform procedural tasks. However limitations of computational tools for automatically identifying and categorising actions in instructions (Van der Sluis et al., 2018, Zhang et al., 2012) still require human intervention as an essential guiding factor. We advocate reader and user studies to explore the relevance of annotation models, to inform further annotation efforts and to inform guidelines for authoring multimodal instructions.

## 5. Acknowledgements

We are grateful for the positive and constructive comments of our ISA reviewers.

## 6. Bibliographical References

- Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard De Melo, and Matthew Stone. 2019. Cite: A corpus of image-text discourse relations. *arXiv preprint arXiv:1904.06286*.
- Yuki M Asano and Gesa Biermann. 2019. Rising adoption and retention of meat-free diets in online recipe data. *Nature Sustainability*, 2(7):621–627.
- Uma Bansal, Aastha Bhardwaj, Som Nath Singh, Sucheta Khubber, Nitya Sharma, and Vasudha Bansal. 2022. Effect of incorporating plant-based quercetin on physicochemical properties, consumer acceptability and sensory profiling of nutrition bars. *Functional Foods in Health and Disease*, 12(3):116–127.
- Roland Barthes. 1977. *Image-music-text*. Macmillan.
- John Bateman, Janina Wildfeuer, and Tuomo Hiippala. 2017. *Multimodality: Foundations, research and analysis—A problem-oriented introduction*. Walter de Gruyter GmbH & Co KG.
- John A Bateman. 2014. Multimodal coherence research and its applications. In *The pragmatics of discourse coherence*, pages 145–177. John Benjamins.
- Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. Recipenlg: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 22–28.
- Skyлар Bowker. 2021. [How to write a recipe post](#).
- Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 35–44.
- Xiaolu Cheng, Shuo-Yu Lin, Kevin Wang, Y Alicia Hong, Xiaoquan Zhao, Dustin Gress, Janusz Wojtusiak, Lawrence J Cheskin, and Hong Xue. 2021. Healthfulness assessment of recipes shared on pinterest: natural language processing and content analysis. *Journal of Medical Internet Research*, 23(4):e25757.
- Kelly Cooper, Ozgur Dedehayir, Carla Riverola, Stephen Harrington, and Elizabeth Alpert. 2022. Exploring consumer perceptions of the value proposition embedded in vegan food products using text analytics. *Sustainability*, 14(4):2075.
- Michelle DiMeo and Sara Pennell. 2018. *Reading and writing recipe books, 1550–1800*. Manchester University Press.
- Myrrh Domingo, Gunther Kress, Rebecca O’Connell, Heather Elliott, Corinne Squire, Carey Jewitt, and Elisabetta Adami. 2014. Development of methodologies for researching online: The case of food blogs.
- David Elsweller, Christoph Trattner, and Morgan Harvey. 2017. Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pages 575–584.
- Janet Floyd and Laurel Forster. 2017. *The Recipe Reader: Narratives-Contexts-Traditions*. Routledge.
- Jill Freyne and Shlomo Berkovsky. 2010. Recommending food: Reasoning on recipes and ingredients. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 381–386. Springer.
- Franck Ganier. 2000. Processing text and pictures in procedural instructions. *Information Design Journal*, 10(2):146–153.

- Franck Ganier. 2012. Cognitive models of processing procedural instructions. *Commun. Technol*, 10:39.
- Manfred Görlach. 1992. Text-types and language history: The cookery recipe. *History of Englishes: New methods and interpretations in historical linguistics*, pages 736–761.
- Kritika Bose Guha and Prakhar Gupta. 2020. Growing trend of veganism in metropolitan cities: Emphasis on baking. *PUSA Journal of Hospitality and Applied Sciences*, 6:22–31.
- Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.
- Carey Jewitt. 2009. *The Routledge handbook of multimodal analysis*, volume 1. Routledge London.
- Pavle Jovanov, Marijana Sakač, Mihaela Jurdana, Zala Jenko Pražnikar, Saša Kenig, Miroslav Hadnadev, Tadeja Jakus, Ana Petelin, Dubravka Škrobot, and Aleksandar Marić. 2021. High-protein bar as a meal replacement in elite sports nutrition: a pilot study. *Foods*, 10(11):2628.
- Mikołaj Kamiński, Karolina Skonieczna-Żydecka, Jan Krzysztof Nowak, and Ewa Stachowska. 2020. Global and local diet popularity rankings, their secular trends, and seasonal variation in google trends data. *Nutrition*, 79:110759.
- Minjeong Kang. 2010. Measuring social media credibility: A study on a measure of blog credibility. *Institute for Public Relations*, 4(4):59–68.
- Joyce Karreman and Nicole Loorbach. 2013. Use and effect of motivational elements in user instructions: What we do and don't know. In *IEEE International Professional Communication 2013 Conference*, pages 1–6. IEEE.
- Joyce Karreman, Nicole Ummelen, and Michaël Steehouder. 2005. Procedural and declarative information in user instructions: What we do and don't know about these information types. In *IPCC 2005. Proceedings. International Professional Communication Conference, 2005.*, pages 328–333. IEEE.
- Gunther R Kress and Theo Van Leeuwen. 2001. Multimodal discourse: The modes and media of contemporary communication. (*No Title*).
- Anna Kustar and Dalia Patino-Echeverri. 2021. A review of environmental life cycle assessments of diets: plant-based solutions are truly sustainable, even in the form of fast foods. *Sustainability*, 13(17):9926.
- Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.
- Guy Lev, Gil Sadeh, Benjamin Klein, and Lior Wolf. 2016. Rnn fisher vectors for action recognition and image annotation. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 833–850. Springer.
- Yiyi Li and Ying Xie. 2020. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of marketing research*, 57(1):1–19.
- Chechen Liao, Pui-Lai To, and Chuang-Chun Liu. 2013. A motivational model of blog usage. *Online Information Review*, 37(4):620–637.
- Leigh Machnee. 2019. Authority, credibility and trust in vegan blogs: Methods used by content creators in the presentation of information. Master's thesis, Department of Computer and Information Sciences, University of Strathclyde.
- Giada Mainolfi, Vittoria Marino, and Riccardo Resciniti. 2022. Not just food: Exploring the influence of food blog engagement on

- intention to taste and to visit. *British Food Journal*, 124(2):430–461.
- Richard E Mayer. 2005. *The Cambridge handbook of multimedia learning*. Cambridge university press.
- Shinsuke Mori, Tetsuro Sasada, Yoko Yamakata, and Koichiro Yoshino. 2012. A machine learning approach to recipe text processing. In *Proceedings of the 1st Cooking with Computer Workshop*, pages 29–34. Citeseer.
- James Pustejovsky. 2018. From actions to events: Communicating through language and gesture. *Interaction Studies*, 19(1-2):289–317.
- James Pustejovsky, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. *Handbook of Linguistic Annotation*, pages 21–72.
- James Pustejovsky, Eben Holderness, Jingxuan Tu, Parker Glenn, Kyeongmin Rim, Kelley Lynch, and Richard Brutti. 2021. Designing multimodal datasets for nlp challenges. *arXiv preprint arXiv:2105.05999*.
- James Pustejovsky and Nikhil Krishnaswamy. 2022. Multimodal semantics for affordances and actions. In *International Conference on Human-Computer Interaction*, pages 137–160. Springer.
- Michaela Regneri, Marcus Rohrbach, Dominikus Wetzels, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.
- Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012a. A database for fine grained activity detection of cooking activities. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1194–1201. IEEE.
- Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012b. Script data for attribute-based recognition of composite activities. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part I 12*, pages 144–157. Springer.
- Markus Rokicki, Christoph Trattner, and Eelco Herder. 2018. The impact of recipe features, social cues and demographics on estimating the healthiness of online recipes. In *Proceedings of the international AAAI conference on web and social media*, volume 12.
- Victoria L Rubin and Elizabeth D Liddy. 2006. Assessing credibility of weblogs. In *AAAI spring symposium: computational approaches to analyzing weblogs*, pages 187–190.
- Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3020–3028.
- Hanna Schösler, Joop De Boer, and Jan J Boersema. 2012. Can we cut out the meat of the dish? constructing consumer-oriented pathways towards meat substitution. *Appetite*, 58(1):39–47.
- Alain D Starke, Martijn C Willemsen, and Christoph Trattner. 2021. Nudging healthy choices in food search through visual attractiveness. *Frontiers in Artificial Intelligence*, 4:621743.
- Christoph Trattner, Dominik Moesslang, and David Elsweiler. 2018. On the predictability of the popularity of online recipes. *EPJ Data Science*, 7(1):1–39.
- Jingxuan Tu, Kyeongmin Rim, and James Pustejovsky. 2022. Competence-based question generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1521–1533.

- Nicole Ummelen. 1997. *Procedural and declarative information in software manuals: Effects on information use, task performance and knowledge*, volume 7. Rodopi.
- Ielka Van der Sluis, Anne Nienke Eppinga, and Gisela Redeker. 2017. Text-picture relations in multimodal instructions. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.
- Ielka Van der Sluis, Lennart Kloppenburg, and Gisela Redeker. 2016a. PAT Workbench: Annotation and evaluation of text and pictures in multimodal instructions. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) at COLING 2016*, pages 131–139.
- Ielka Van der Sluis, Shadira Leito, and Gisela Redeker. 2016b. Text-picture relations in cooking instructions. In *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation: Proceedings of the Twelfth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 22–27.
- Ielka Van der Sluis and Hanna Mellema. Submitted. A recipe for success: The design, use and effectiveness of multimodal online baking instructions. *Multimodality & Society*.
- Ielka Van der Sluis and Gisela Redeker. 2019. The pat annotation model for multimodal instructions. In *6th European and 9th Nordic Symposium on Multimodal Communication*.
- Ielka Van der Sluis, Gisela Redeker, and Sannah Debreczeni. 2022. A text-based method to derive the main action structure in procedural instructions. In *AREA II: Workshop on the Annotation, Recognition and Evaluation of Actions held in conjunction with the 33rd European Summer School in Logic, Language and Information 8-19 August, 2022*.
- Ielka Van der Sluis, Renate Vergeer, and Gisela Redeker. 2018. Action categorisation in multimodal instructions. In *Proceedings of (AREA 2018)*, pages 22–27.
- Youri Van Pinxteren, Gijs Geleijnse, and Paul Kamsteeg. 2011. Deriving a recipe similarity measure for recommending healthful meals. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 105–114.
- Charlotte Vijfvinkel, Ielka Van der Sluis, and Gisela Redeker. 2018. I like to move it move it: Analysing first-aid instruction videos for moving a victim. In *TABU Dag 2018: The 39th International Linguistics Conference*.
- Janina Wildfeuer, Ielka Van der Sluis, Gisela Redeker, and Nina Van der Velden. 2022. No laughing matter!? analyzing the page layout of instruction comics. *Journal of Graphic Novels and Comics*, pages 1–22.
- Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*.
- Asami Yajima and Ichiro Kobayashi. 2009. "easy" cooking recipe recommendation considering user's conditions. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 13–16. IEEE.
- Yu Zhang, Li Cheng, Jianxin Wu, Jianfei Cai, Minh N Do, and Jiangbo Lu. 2016. Action recognition in still images with minimum annotation efforts. *IEEE Transactions on Image Processing*, 25(11):5479–5490.
- Ziqi Zhang, Philip Webster, Victoria S Uren, Andrea Varga, and Fabio Ciravegna. 2012. Automatically extracting procedural knowledge from instructional texts using natural language processing. In *LREC*, volume 2012, pages 520–527. Citeseer.

# Author Index

- Aktas, Berfin, [60](#)  
Alcaina, Cristina Fernández, [66](#)  
Amorim, Evelin, [99](#)  
Anisimova, Mariia, [17](#)
- Bäckström, Linnéa, [93](#)  
Böbel, Nina, [93](#)  
Bunt, Harry, [111](#), [122](#)
- Croft, William, [93](#)
- Dong, Min, [144](#)
- Edlund, Jens, [1](#)  
Er, Mustafa Erolcan, [53](#)  
Esfandiari-Baiat, Ghazaleh, [1](#)
- Fang, Alex Chengyu, [144](#)  
Fu, Yingxue, [27](#)  
Fučíková, Eva, [66](#)
- Hajič, Jan, [66](#)  
Haller, Armin, [71](#)
- Jezek, Elisabetta, [47](#)
- Kiewiet de Jonge, Jarred, [152](#)  
Kurfalı, Murathan, [53](#)
- Leal, António, [99](#)  
Lee, Kiyong, [133](#)  
Ljunglöf, Peter, [93](#)  
Lorenzi, Arthur, [93](#)  
Lyngfelt, Ben, [93](#)
- Malchanau, Andrei, [122](#)  
Marini, Costanza, [47](#)  
Matos, Ely E., [93](#)  
Meeus, Quentin, [8](#)  
Moens, Marie-Francine, [8](#)
- Naseem, Usman, [71](#)
- Özmen, Burak, [60](#)
- Paccosi, Teresa, [39](#)  
Petliak, Nataliia, [66](#)
- Petukhova, Volha, [122](#)
- Rodriguez Mendez, Sergio J., [71](#)
- Salman, Muhammad, [71](#)  
Silvano, Purificação, [99](#)
- Timponi Torrent, Tiago, [93](#)  
Tomaszewska, Aleksandra, [99](#)  
Tonelli, Sara, [39](#)
- Uhrig, Peter, [93](#)  
Urešová, Zdeňka, [66](#)
- Van de Cruys, Tim, [82](#)  
van der Sluis, Ielka, [152](#)  
Van hamme, Hugo, [8](#)  
Vanroy, Bram, [82](#)
- Zeng, Jiamei, [144](#)  
Zeyrek, Deniz, [53](#)  
Ziem, Alexander, [93](#)  
Zikánová, Šárka, [17](#)