The Workshop on Cognitive Aspects of the Lexicon (CogALex@LREC-COLING 2024)

Workshop Proceedings

Editors
Michael Zock, Emmanuele Chersoni, Yu-Yin Hsu and Simon De Deyne

20 May, 2024
Torino, Italia

Proceedings of the Workshop on Cognitive Aspects of the Lexicon (CogaLex@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 978-2-493814-45-6
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

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

Message from the Program Chairs

Words are the exchange money for revealing our thoughts. While a single word may capture a world of knowledge in a nutshell, its combination with other words may allow for even more, support communication, memorization, and even thinking, an activity that consists largely of perceiving links, and combining concepts for which we have created words

Words can entertain, reveal, or hide. Sharper than razor blades, they can also make us laugh or cry. No doubt, words are important. They are everywhere and multifarious, and different disciplines have contributed to explain how they are learned, represented, organized, or used. While we must continue to draw on expertise from various fields (linguistics, psychology, etc.), it is essential to facilitate the communication of the most recent developments in each of them. For example, widely used techniques like association thesauri, complex graphs, or large language models (LLMs) have not been created by linguists but by psychologists or specialists in AI (mathematicians, statisticians, and computer scientists). The knowledge in each of these areas is growing at an exceptional rate, and what used to be true yesterday may not be so anymore today or tomorrow. This workshop aims to take stock of these developments and support the continuing exchange between researchers across disciplines. These are but some of the motivations for organizing this kind of workshop. For more details concerning the landscape, the diversity of problems and solutions, see our homepage: `https://sites.google.com/view/cogalex-viii-2024/home`.

This is the 8th CogALex workshop. Following the precursor 'Enhancing and Using Electronic Dictionaries' (2004, COLING, Geneva, Switzerland), there have been seven similar events, though with a stronger focus on the 'Cognitive Aspects of the Lexicon', hence the name and acronym of this workshop, CogALex:

- CogALex-I (2008, COLING, Manchester, UK)
- CogALex-II (2010, COLING, Beijing, China)
- CogALex-III (2012, COLING, Mumbai, India)
- CogALex-IV (2014, COLING, Dublin, Ireland)
- CogALex-V (2016, COLING, Osaka, Japan)
- CogALex-VI (2020, COLING, Barcelona, Spain)
- CogALex-VII (2022, AACL-IJCNLP, Taipei, Taiwan)

As every organizer knows, no workshop can be held without the help of competent and devoted reviewers (`https://sites.google.com/view/cogalex-viii-2024/home/programme-committee`). We would like to take this opportunity to thank all of them for their hard work. Last, but not least we would like to thank Gilles-Maurice de Schryver who was so kind to accept our invitation as an invited speaker. Being an experienced professional, we are particularly keen to hear his thoughts about the potential uses of LLMs for dictionary making (`https://sites.google.com/view/cogalex-viii-2024/home/invited-speaker`).

In this edition of the workshop, we have received 24 submissions and accepted 19 of them for presentation (79% of acceptance rate).

One last point to the authors. We have received an unusually high number of excellent papers, which would easily fit a full day workshop. Unfortunately only half a day workshop could be scheduled this year, which means that we could not have enough oral presentation slots available reflecting the high quality of submissions. We have tried to balance the time for oral and poster presentations to do justice to both. We'd like to thank all authors for submitting their best work and look forward to meeting you in Turin.

The Organizers
Michael Zock, Emmanuele Chersoni, Yu-Yin Hsu & Simon De Deyne

Organizing Committee

Michael Zock (CNRS & Aix-Marseille Université, France)
Emmanuele Chersoni (The Hong Kong Polytechnic University, Hong Kong, China)
Simon De Deyne (University of Melbourne, Australia)
Yu-Yin Hsu (The Hong Kong Polytechnic University, Hong Kong, China)

Program Committee

Verginica Barbu Mititelu (RACAI, Romania)
Chris Biemann (University of Hamburg, Germany)
Jose Camacho-Collados (Cardiff University, United Kingdom))
Ya-Ning Chang (University of Cambridge, United Kingdom)
Kenneth Church (Northeastern University, United States)
Rodolfo Delmonte (Ca' Foscari University of Venice, Italy)
Olivier Ferret (Institut CEA List, France)
Thierry Fontenelle (European Investment Bank, Luxembourg)
Thomas François (Université Catholique de Louvain, Belgium)
Markus J. Hofmann (Bergische Universität Wuppertal, Germany)
Shu-Kai Hsieh (National Taiwan University, Taiwan)
Ignacio Iacobacci (Huawei Noah's Ark Lab, United Kingdom)
Mathieu Lafourcade (Université de Montpellier, France)
Phillippe Langlais (Université de Montréal, Canada)
Michael S. Vitevitch (University of Kansas, United States)
Adam Pease (Articulate Software, United States)
Alessandro Raganato (University of Milano-Bicocca, Italy)
Giulia Rambelli (University of Bologna, Italy)
Didier Schwab (Université Grenoble Alpes, France)
Cynthia Siew (National University of Singapore, Singapore)
Massimo Stella (University of Trento, Italy)
Gilles Sérasset (Université Grenoble Alpes, France)
Carole Tiberius (Leiden University, Netherlands)
Yu Wang (The Hong Kong Polytechnic University, Hong Kong, China)
Werner Winiwarter (University of Vienna, Austria)
Pierre Zweigenbaum (Université Paris-Saclay, France)

Table of Contents

Workshop Program

Monday, May 20, 2024

9:00–9:15 *Introduction*
Michael Zock

9:15–10:15 Keynote Speech

9:15–10:15 *Fine-tuning LLMs for lexicography*
Gilles-Maurice de Schryver

10:15–10:40 Oral Presentations 1

CLAVELL - Cognitive Linguistic Annotation and Visualization Environment for Language Learning
Werner Winiwarter

10:40–11:00 Coffee Break

11:00–12:00 Poster Session

Individual Text Corpora Predict Openness, Interests, Knowledge and Level of Education
Markus J. Hofmann, Markus T. Jansen, Christoph Wigbels, Benny Briesemeister and Arthur M. Jacobs

An Empirical Study on Vague Deictic Temporal Adverbials
Svenja Kenneweg, Brendan Balcerak Jackson, Joerg Deigmoeller, Julian Eggert and Philipp Cimiano

Symbolic Learning of Rules for Semantic Relation Types Identification in French Genitive Postnominal Prepositional Phrases
Hani Guenoune and Mathieu Lafourcade

How Human-Like Are Word Associations in Generative Models? An Experiment in Slovene
Špela Vintar, Mojca Brglez and Aleš Žagar

Idiom Complexity in Apple-Pie Order: The Disentanglement of Decomposability and Transparency
Irene Pagliai

Monday, May 20, 2024 (continued)

12:00–12:50	**Oral Presentations 2**
	Combining Neo-Structuralist and Cognitive Approaches to Semantics to Build Wordnets for Ancient Languages: Challenges and Perspectives Erica Biagetti, Martina Giuliani, Silvia Zampetta, Silvia Luraghi and Chiara Zanchi
	SensoryT5: Infusing Sensorimotor Norms into T5 for Enhanced Fine-grained Emotion Classification Yuhan Xia, Qingqing Zhao, Yunfei Long, Ge Xu and Jia Wang
12:50–13:00	*Conclusion* Michael Zock

CLAVELL – A Cognitive Linguistic Annotation and Visualization Environment for Language Learning

Werner Winiwarter
University of Vienna, Faculty of Computer Science
Währingerstrasse 29, 1090 Vienna, Austria
werner.winiwarter@univie.ac.at

Abstract

In this paper, we introduce a novel sentence annotation method based on Radical Construction Grammar and Uniform Meaning Representation, covering multiple levels of linguistic analysis, ranging from interlinear morphemic glossing to PropBank rolesets, WordNet synsets, and Wikipedia page titles as concept identifiers. We visually enhance our annotation by using images to represent concepts, emojis for roles, and color-coding for constructions, the fundamental concept of Construction Grammar. The meaning representation is embedded into the syntactic parse by aligning all concepts with the surface tokens in the sentence. The main motivation for developing this type of representation was its use in second language acquisition as part of a Web-based language learning environment. By engaging in entertaining annotation tasks students assemble incrementally the representation using a bottom-up strategy. Based on language exposure while performing these exercises, we populate personal idiolectal constructicons representing the students' current status of second language comprehension. To showcase our system, we have implemented it for Japanese because of its soaring popularity in our language education program and the particular problems it poses to those trying to learn this language, especially Westerners.

Keywords: cognitive linguistics, Radical Construction Grammar, Uniform Meaning Representation, visual annotation, second language acquisition, Web-based language learning, constructicon, Japanese linguistics

1. Introduction

In the shadow of the latest LLM hype, there have been nonetheless significant developments in the field of cognitive linguistics. Two remarkable recent achievements have been Radical Construction Grammar (RCG) and Uniform Meaning Representation (UMR), which pave the way for cross-linguistic semantic annotation of documents. Building on this research work, we have extended the representation towards an interlingual annotation by linking all concepts to knowledge bases, mapping PropBank core arguments to interpretable roles, aligning concepts with words, and enhancing the display with visual elements.

Language understanding is a multi-step process where a signal is broken down into smaller units, e.g. words, morphemes, sounds, or letters, to be then interpreted in terms of meaning. Put differently, to understand, we need to recognize the categories standing for meaning or meaning relations. Categorization is a fundamental process needed both to understand the meaning of a sentence as well as to understand the regularities of the mappings between "meanings" and "forms". Thus, students have to learn two kinds of language, the target language, i.e. the one they are exposed to, and a meta-language allowing them to describe regularities.

Language learning is already difficult but it may well be even more of a challenge as one attempts to learn a typologically different language. For example, a European trying to learn Chinese has to attune his ears as Chinese has not only many unknown sounds but also specific intonations (tones) for each "word". Next to mastering the sound system, Chinese is even more intimidating when it comes to reading and writing. Regarding this aspect, things are even worse for Japanese, which has three writing systems: two for the sound (hiragana and katakana) and one mixed for meaning and sound (kanji). Drawing heavily on Chinese characters, kanji encode meaning and sound, though not always in a very regular form. To make things even more complicated, kanji most often have at least two kinds of pronunciation, one based on Chinese, the other based on Japanese. All these are difficult hurdles for the learner of Japanese. As we have already addressed many of these issues in our previous work (Wloka and Winiwarter, 2021a,b; Winiwarter and Wloka, 2022), we will focus here only on using our annotation for Japanese language learning. The goal is to alleviate the burden of converting forms to meaning and to ease the gaining of certain insights concerning the mechanics and functioning of language.

Due to the worldwide manga craze there has been an unprecedented increase in demand for Japanese language courses. As a result, there have been requests from our language education department for technological support. Our Web-based annotation exercises are supposed to im-

The Workshop on Cognitive Aspects of the Lexicon, pages 1–13
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

prove language learning in an appealing way. One central component of our environment are personal constructicons, which reflect each student's progress, proficiency, and individual learning path. Aggregating this knowledge is very useful as it offers invaluable insights for instructors and course designers alike.

This paper is organized as follows. In section 2, we provide some background information, citing relevant related work. In section 3, we discuss the representation of major propositional acts in RCG. In section 4, we describe implementation details, and finally, in section 5, we provide an outlook towards future work.

2. Background and Related Work

2.1. Construction Grammar

Construction Grammar (Hoffmann and Trousdale, 2013; Hilpert, 2014; Ungerer and Hartmann, 2023) is actually a broad family of theories in the area of *cognitive linguistics* (Croft and Cruse, 2004). Their common denominator is the view that *constructions* are the basic units of language, which are pairings of *form* and *meaning*. Researchers working in this paradigm/framework reject the separation of the grammar and the lexicon. Both are considered to be constructions.

Radical Construction Grammar (Croft, 2001, 2022) (RCG) was designed with *typological* applicability as main motivation. Linguistic *typology* (Croft, 2002) studies and classifies languages according to their structural features to allow their comparison. RCG considers word classes and other syntactic structures as language-specific and construction-specific (Croft, 2023). In a recent supplement to (Croft, 2022), the author provides a taxonomic/partonomic tree of constructions[1], therefore supporting an ontological view on constructions as a structured set of concepts, which, if properly combined represent the meaning underlying a sentence. For more background on this line of thought we refer to (Zock et al., 2008; Borgo et al., 2019).

One of the main consequences of Construction Grammars is that the traditional divide between lexicon and grammar is abandoned. Everything is a construction. For example, the lexicon is replaced by a network of constructions, words being constructions. Most attempts to assemble such a *constructicon* are in the context of the *FrameNet*[2] project (Lyngfelt, 2018). Whereas the original model was a *taxonomical inheritance network*, recent research on *usage-based linguistics* (Divjak, 2019) has led to a *multidimensional association network* (Diessel, 2023). This is in line with most of the work done on lexical graphs: WordNet (Fellbaum, 1998; Miller, 1990), FrameNet (Fillmore et al., 2003), BabelNet (Navigli and Ponzetto, 2012), VerbNet (Kipper-Schuler, 2005), MindNet (Richardson et al., 1998), HOWNET (Dong and Dong, 2006), ConceptNet (Liu and Singh, 2004; Speer et al., 2017), YAGO (Suchanek et al., 2007), DBnary (Sérasset, 2015), or JeuxDeMots (Lafourcade et al., 2015). Recently, related experiments were conducted to explore the role of *language exposure* on emergence and fading of constructions in the constructicon (Dunn, 2022). This novel interpretation of constructicons is also highly compatible with most modern views on the mental lexicon (see Papafragou et al., 2022; Zock and Biemann, 2020; Zock, 2022).

2.2. Meaning Representations

The annotation of sentences with *meaning representations* has established itself in the last decade as a thriving research field in computational linguistics (see Abend and Rappoport, 2017). The most influential and most actively promoted approach has been the *Abstract Meaning Representation*[3] (AMR) (Banarescu et al., 2013). There are many parsers available, the best[4] one being at the moment Lee et al. (2022). The *SPRING* parser can be tried via a Web interface[5], which also offers a nice visualization. One point of criticism concerning AMR is its reliance on numbered, hence not directly interpretable core arguments; it is addressed by the *WISeR* meaning representation (Feng et al., 2023a), which maps them to thematic roles.

AMR has been recently extended to the *Uniform Meaning Representation*[6] (UMR) (Gysel et al., 2021). It enhances AMR by adding support for other languages (in particular low-resource languages), and a document-level representation capturing intersentential coreference and temporal/modal dependencies. There is an upcoming workshop to kick-start the development of UMR parsers. According to the *UMR guidelines*[7], UMR fully embraces RCG as a theoretical foundation.

2.3. Interlinear Morphemic Glossing

Whereas this topic has been neglected for a long time by natural language processing research, it has a long tradition in linguistics and typology. An *interlinear morphemic gloss* (IMG) represents a text

[1] https://www.unm.edu/~wcroft/Papers/ConstructionRelations.pdf

[2] https://framenet.icsi.berkeley.edu/

[3] https://amr.isi.edu/

[4] https://paperswithcode.com/task/amr-parsing/latest

[5] http://nlp.uniroma1.it/spring/

[6] https://umr4nlp.github.io/web/

[7] https://github.com/umr4nlp/umr-guidelines/

as a string of elements by ideally mapping each morpheme of the source language to a morpheme of the target language or to a grammatical category.

Until the recent past (Leipzig Glossing Rules[8]), there was no common format, which resulted in a confusing variety of glossing styles. This has changed. Recently there have been some efforts by computational linguists to extend and formalize the guidelines (Mortensen et al., 2023) and to automate interlinear glossing (e.g. Samardžić et al., 2015; Zhao et al., 2020; Barriga Martínez et al., 2021). There has even been a first shared task on this topic in 2023 (Ginn et al., 2023).

2.4. Multimodal Resources

Multimodal enhancements of lexical resources have a long history but only recently they have gained momentum due to the interest on *visual question answering* (Lerner et al., 2024) or *multimodal large language models* (Bewersdorff et al., 2024). One example of an attempt towards a multimodal semantic representation is *VoxML* (Pustejovsky et al., 2016).

WordNet has been extended by *ImageNet*, which maps about 1,000 images to each synset (Deng et al., 2009). Another effort to assign cliparts to a small set of synsets was proposed by (Bond et al., 2009). Alas, this project is discontinued. A much more influential resource is *Wikipedia*, which has been increasingly enhanced with visual representations. However, the number of images provided varies widely depending on the language.

The most comprehensive effort is *BabelNet*[9] (Navigli et al., 2021) with the annotation tool *Babelfy*[10] (Moro et al., 2014) and the latest *BabelPic*[11] (Calabrese et al., 2020) dataset targeting abstract concepts.

Even though the use of pictorial illustrations has a long history in language teaching textbooks, there is a crying need for visual representations of meaning representations of sentences.

2.5. Japanese Language

Japanese is an agglutinative SOV language with topic-comment sentence structure. Both agglutinative languages and fusional languages like, for instance, German, are synthetic languages, i.e. they are statistically characterized by a higher morpheme-to-word ratio. In agglutinative languages, words contain multiple morphemes concatenated together in such a manner that individual word stems and affixes can be usually isolated and identified, whereas fusional languages combine multiple grammatical categories into one affix. Therefore, agglutinative languages tend to have more easily deducible word meanings compared to fusional languages, which allow unpredictable modifications in either or both the phonetics or spelling of one or more morphemes within a word.

In Japanese, phrases are exclusively head-final and compound sentences are strictly left-branching. The most noticeable characteristics for foreigners are the lack of articles (a/the), the absence of markers for number (sg. vs pl.) or gender (masculine/feminine), or the fact that adjectives are conjugated. On the other hand, Japanese has a complex system of honorifics, high dependency on context, hence strong likelihood of ambiguity, due to the omission of the subject or the use of zero anaphora.

There are many excellent reference grammars, e.g. Bowring and Laurie (1992); Kamermans (2010); Kaiser et al. (2013), and a lot of research done by Japanese linguists: see Hasegawa (2015) for an introduction; for a recent comprehensive overview we recommend Hasegawa (2018). There is also a wealth of typological studies of Japanese, e.g. Takezawa (1993); Washio (1997); Matsumoto (1997); Taoka (2000); Ohori (2001); Yuasa and Sadock (2002); Iwasaki (2013). Based on the *Japanese FrameNet*[12] project (Ohara et al., 2003), there have been ongoing efforts towards a Japanese constructicon (Ohara, 2014). There is also an important lexicographic work by Backhouse (2016), who organizes Japanese vocabulary according to semantic frames.

One of the main obstacles for getting proficient in Japanese is the complex writing system (see Matsumoto, 2007; Mori, 2014; Paxton, 2019). It uses a combination of logographic *kanji* and two syllabaries *hiragana* and *katakana*. Kanji are adopted Chinese characters. Since 2010 Japanese students are required to learn 2,136 so-called *jōyō kanji* in primary and secondary school.

There exist several *romanization* systems, i.e. using Latin script to write Japanese. The most widely used one is the *Hepburn romanization*, which has several variants, the most common one being the *Revised Hepburn* (see Kudo, 2011). There are many romanization tools, the most easily accessible one for the use in natural language processing software written in Python is *Pykakasi*[13] based on the *kakasi*[14] library.

The most important lexical resource for Japanese is the *Japanese Multilingual dictionary* (JMdict) (Breen, 2004), which can be searched

[8] https://www.eva.mpg.de/lingua/tools-at-lingboard/glossing_rules.php
[9] https://babelnet.org/
[10] http://babelfy.org/
[11] https://sapienzanlp.github.io/babelpic/
[12] https://jfn.st.hc.keio.ac.jp/
[13] https://pypi.org/project/pykakasi/
[14] http://kakasi.namazu.org/

online in combination with many other lexical resources via the *Online Japanese Dictionary Service* (WWWJDIC)[15]. Another very useful online service is *Honyaku Star*[16]. It references numerous dictionaries and corpora and shows translations in context. Honyaku Star includes currently over 2 million translations.

Japanese is also part of the *Open Multilingual Wordnet* (Bond and Paik, 2012)[17], which allows the mapping of Japanese words to English synsets. It is easily accessible via the *NLTK* toolkit[18].

The most prolific linguistic tool for Japanese is certainly the *CaboCha* dependency parser (Kudo and Matsumoto, 2002), which includes the *MeCab* part-of-speech and morphological analyzer (Kudo et al., 2004). More recently, trained pipelines have been added to the popular natural language toolkit *SpaCy*[19]. Another similar solution is *UniDic2UD*[20].

3. Meaning Representation in RCG

In RCG there are two central *comparative concepts* (see Haspelmath, 2010), i.e. theoretical concepts used for crosslinguistic comparison. The first one is the **construction (cxn)**, which is defined as any pairing of form and function in any language to express a particular combination of semantic content and information packaging (see Croft, 2022). The second comparative concept is the **strategy**, which further distinguishes certain characteristics of grammatical form defined in a crosslinguistically consistent way.

There are three fundamental **information packaging** functions that structure phrases and clauses: reference, modification, and predication. They are called **propositional act** functions and correspond to the prototypical **semantic classes**: objects, properties, and actions. However, any semantic class can be packaged in any information packaging function so that we end up with a 3×3 matrix. In the following subsections we provide an example for each cell of this matrix by introducing our annotation for the resulting constructions.

All theoretic concepts defined in (Croft, 2022) are emphasized in boldface. We provide a short definition for each term, for a more detailed description with examples we refer to the voluminous glossary of (Croft, 2022). As much as possible, RCG relies on terms already in use in linguistics, e.g. *reference* or *topic*, and while they try to make their definitions more precise, quite so often they depart from the traditional use.

3.1. Reference

The information packaging function **reference** indicates what the speaker is talking about. The prototypical semantic class are **object** concepts, which include persons, animals, and physical objects.

In Fig. 1, the first basic construction is an example of **object reference**. We annotate the original orthographic representation from the source text with the following information:

- morphemic representation,
- interlinear morphemic gloss,
- translation to concept(s),
- visualization of concept(s),
- construction label.

Figure 1: Three examples of reference.

We use Revised Hepburn romanization for the morphemic representation with some additional information. For example, the capitalized reading "TŌ" indicates that this is a Sino-Japanese reading. It is translated to the WordNet synset `tower.n.01`. By clicking on the image, an enlarged version can be inspected including the synset gloss as caption. The resulting construction is a **referent expression**, i.e. its prototypical use would be as **head** of a

[15] http://wwwjdic.se/
[16] http://honyakustar.com/
[17] https://omwn.org/
[18] https://www.nltk.org/
[19] https://spacy.io/models/ja
[20] https://github.com/KoichiYasuoka/UniDic2UD

referring phrase. Both semantic content and information packaging are color-coded in the annotation. The former as background color of the image, the latter as border color. As this is the prototypical combination, both are drawn in magenta.

The second construction in Fig.1 packages a **property** as reference. Properties are relational, 1-dimensional, usually scalar and stable concepts, which are drawn in green. In this example, the property "utsukushi-" ("beautiful") is translated to the PropBank roleset `beautiful-02`. The suffix "sa" acts as *nominalizer* (NR) to derive the referent expression "beauty".

By analogy, the third construction represents an **action**, i.e. relational, dynamic, and transitory concepts painted in cyan. The action "nusu-" ("steal") is translated to the PropBank roleset `thieve-01`. The continuative ending "mi" again acts as *nominalizer* (NR) resulting in the referent expression "theft".

3.2. Modification

The second propositional act provides additional information about the referent and enriches the specification of the referent for the hearer. The prototypical construction is **property modification** and is shown in Fig. 2. The property "cute" is a **modifier** for the referent expression "dog" (border color green), i.e. the **head** of an **attributive phrase**. By combining the two elements, we get a **modification cxn**, which is defined as a **referring phrase** (color magenta) consisting of a referent expression and one or several attributive phrases.

The grammatical category NPST in the gloss for "kawai-i" indicates the tense non-past, because "kawai-i" is an "i-adjective" that behaves like a verb (or, conceptually speaking, plays the same role as a verb). Therefore, it is also referred to as "verbal adjective" in many reference grammars.

In the meaning representation, the two concepts are linked by a relation with the role MOD (for more details on roles in UMR we refer to[21]). As mentioned before, we use emojis for the roles, in this case a ribbon 🎀. For an optimal alignment with the structural representation, the left-right axis conveys meaning. We do not add arrowheads to the relations because almost all relations in our annotation point from right to left due to the left-branching nature of Japanese language, therefore we define this direction as default interpretation.

Figure 3 shows an example of **object modification**. The postposition "no", indicating the modificaton relation, is not annotated. We allow to omit annotations for frequent monosyllabic postpositions in

[21] https://github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md

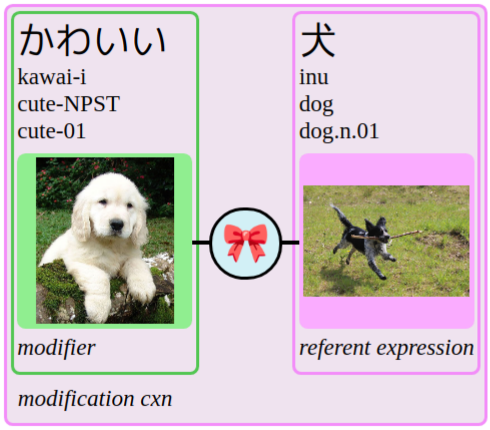


Figure 2: Example of property modification.

language learning scenarios because of their abundant use in Japanese and their excessive polysemy and homonymy.

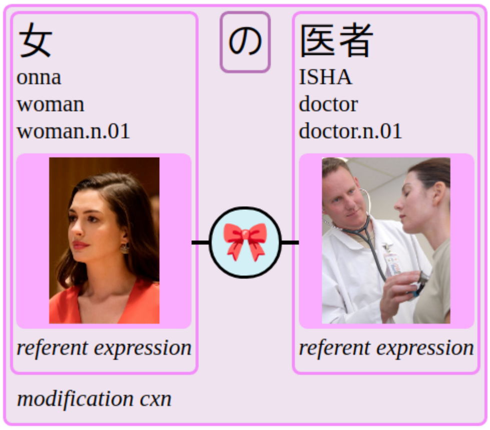


Figure 3: Example of object modification.

The third combination, **action modification**, is realized in Japanese as **relative clause**, which is modeled in RCG together with a **matrix clause** as modifier of the head of the referring phrase, the **relative clause head**, to result in a **relative clause cxn**. In the example in Fig. 4, only the relative clause and the relative clause head are shown.

The action "lose" is packaged as a **predicate**, the **head** of a **clause**. The past tense (PST) is modeled in the meaning representation as a BEFORE (⏪) relation to the special concept *DCT (document creation time)*, colored in orange, representing the present moment. This sequence of concepts is indicated in the gloss by a vertical bar. The information-packaging alternation passive voice (PASS) is not reflected in the meaning representation. Finally, the grave accent in the morphemic representation indicates that the native reading of the second kanji is unvoiced, i.e. "kami" and only changes to "gami" as second part of “tegami”. This voicing is called *rendaku* in Japanese.

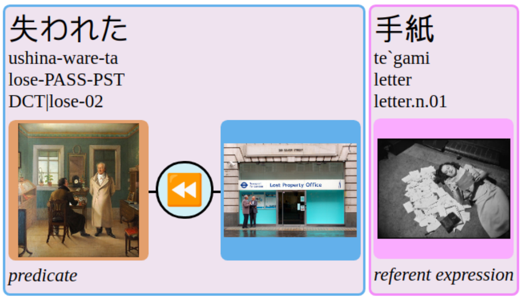


Figure 4: Example of action modification.

3.3. Predication

The third propositional act conveys what the speaker is asserting about the referents in a particular utterance. As we have already given an example of the prototypical construction **action predication** as part of Fig. 4, we focus on the two non-prototypical constructions.

An example of **object predication** can be seen in Fig. 5. The postposition “wa” indicates the **topic** of the sentence, i.e. the referent in a **topic–comment** information packaging that the **comment** is predicated about. In the meaning representation, the object predication is modeled by a special concept, which is aligned with the **copula** (COP) “da”. The predication asserts what object CATEGORY (🏷️) the THEME (📜) belongs to.

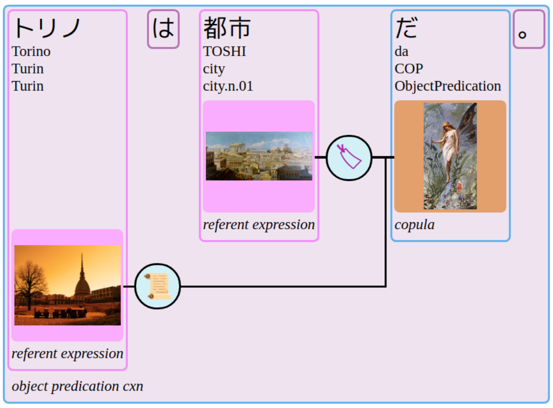


Figure 5: Example of object predication.

Finally, Fig. 6 shows an example of **property predication**. Since “kashiko-i” is also a “verbal adjective” just like “kawai-i” in Fig. 2, there is no copula and the special concept for the property predication is directly appended to the property concept in the meaning representation and linked to it by a PROPERTY (🌈) relation.

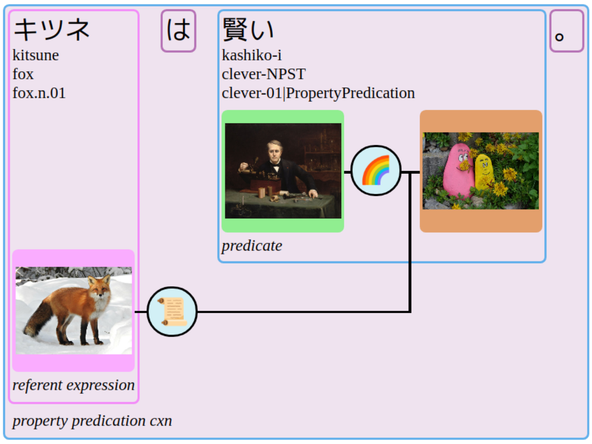


Figure 6: Example of property predication.

4. Implementation

Figure 7 highlights the main components of our system architecture. We have realized our language learning environment as Web-based solution, so that the language students can access the learning server through a Web browser by using augmented browsing enabled through *Chrome extension APIs*[22], and the *jQuery*[23] and *jQuery UI*[24] libraries. Whenever a student loads a new Japanese Web document, it is automatically analyzed and segmented into individual sentences. Each sentence is augmented with an event handler. If a student then clicks on a sentence, it is transferred to the server via *XMLHttpRequests*.

The language learning server is implemented in *SWI-Prolog*[25] (Wielemaker et al., 2012), which is not only predestined for natural language tasks but also provides a scalable Web server solution (Wielemaker et al., 2008) and libraries for efficiently handling RDF and XML files.

The server parses the sentence by using the linguistic knowledge stored in the *personal idiolectal constructicon* (reflecting the student’s unique use of the learned language based on past exposure) and dynamically generates an HTML page with the annotated sentence, which is opened in a new tab in the student’s browser.

The user can now add new information to the annotation, which is again sent to the server leading to an update of the constructicon, a reparsing of the sentence, and an actualization of the HTML page.

As external resources we use *PropBank Frame Files* (Pradhan et al., 2022), *WordNet* (Princeton University, 2012), *DBpedia* (Lehmann et al.,

[22] https://developer.chrome.com/docs/extensions/reference/api
[23] https://jquery.com/
[24] https://jqueryui.com/
[25] https://www.swi-prolog.org

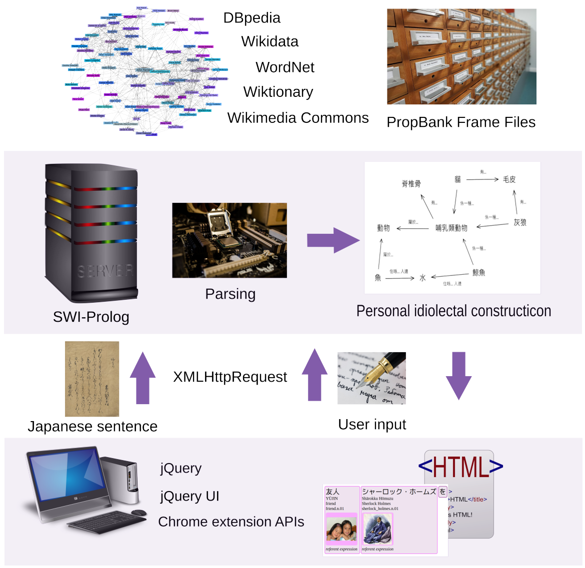


Figure 7: System architecture.

2015), *Wikidata*[26], *Wiktionary*[27], and *Wikimedia Commons*[28].

In Fig. 8, we take a closer look at the user interaction. In order to populate their personal idiolectal constructicons, the language students start at the beginning of their training from scratch with an empty knowledge base.

We have designed the annotation tasks with gaming elements in mind in that we first confront the students with a list of Japanese characters, which is the starting point to assemble the complete annotation step-by-step following a bottom-up strategy.

As Step 1 in Fig. 8, the user can form words by drawing a box with the cursor to select characters. Selected characters are highlighted in orange. As soon as the user releases the mouse button, the display changes to Step 2.

Now, the student is supposed to enter the correct morphemic representation, the interlinear morphemic gloss, and the translation to concept(s). At every step, i.e. for every input, the level of support offered to the user can be increased by displaying select menus or suggestions. This is achieved by accessing the external resources in Fig. 7 as well as the language-specific tools and lexicons mentioned in section 2.5.

As soon as the user adds some information, it is stored in the constructicon, and can be used to learn rules to apply this linguistic knowledge to new examples. Context-sensitive rules are learnt automatically and adjusted incrementally for each new item. We also store the number of times the student was exposed to this item so we can choose

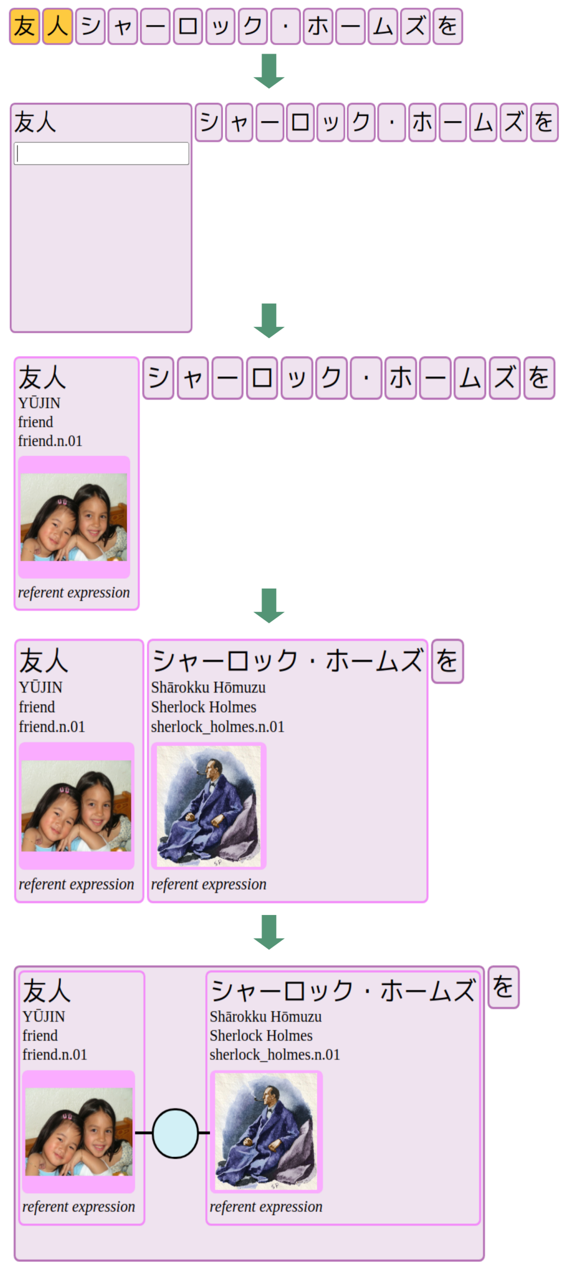


Figure 8: Example of user interaction.

the number of repetitions the student has to perform before the item is inserted automatically.

The information in the constructicon is organized as associative network, which is stored persistently using Prolog fact files. In addition, we offer routines to aggregate and focus on certain aspects like words, phrasemes, or production rules to export them in common exchange formats.

After entering all the data, the complete basic construction including the visualization of the concept(s) is displayed in Step 3. We offer default images for all concepts, which are taken from Wikimedia Commons. Whenever possible we automatically extract the image from the corresponding

[26] https://www.wikidata.org/
[27] https://en.wiktionary.org/
[28] https://commons.wikimedia.org/

Wikipedia page using DBpedia as well as Wikidata and Wiktionary for retrieval. However, since some associations between concepts and images can be culture-specific, the instructors can freely localize the images to adapt them to their target audiences. In addition, we encourage the students to choose their own images to personalize the learning experience and to collect valuable data about the observed multimodal associations.

In Step 4, the user has completed the annotation of the first two constructions. By selecting both of them with the mouse, we end up in Step 5, i.e. with a new complex construction and a new relation between the two concepts. The last thing the student has to do is to input the type of the new construction and the role of the new relation. Again, assistance can be offered for these two annotation steps. Also, the emojis for the different roles can be altered to suit the personal preferences of the student.

Figure 9 displays the complete annotation for the whole sentence. It is the first sentence from the Japanese translation of the Sherlock Holmes short story "The red-headed league" by Sir Arthur Conan Doyle. This translation is available at the Japanese digital library *Aozora Bunko*[29] under a Creative Commons license. Aozora Bunko contains over 17,000 literary works without copyright and therefore represents an invaluable resource for Japanese language education, literary studies, as well as translation studies.

It has to be emphasized that the visualization of the annotation is realized purely by using HTML and CSS without any additional libraries apart from basic jQuery UI widgets. This way we can gain flexibility towards alternative application scenarios which would favor a representation which is directly embedded into the original Web page, by avoiding conflicts with original libraries and control flow. One consequence of this design choice is that we only include concepts and roles in the Web document generated at the language server and then add the connecting lines for the relations dynamically at the client after loading the page for an exact rendering of start and end points.

The only new constructions in Fig. 9 not covered so far are the **auxiliary** "koto ga atta", an expression, which indicates the perfect (PRF) ASPECT (💅), the corresponding **auxiliary cxn**, and **clause** for the complete sentence.

Additional roles are TEMPORAL (⏰), SEASON (🌻), and UNDERGOER (🚑). The background color green for the role 🌻 indicates an *inverse relation*, i.e. "autumn" is a SEASON-OF "last year". Inverse relations are mainly used for focusing in UMR, in this case on "autumn". A similar situation is the relation between the perfect aspect and the predicate "visited". However, in this case the predicate is the head of the auxiliary cxn, therefore we change the direction of the relation from left to right, indicated by the line color violet. Finally, a hopefully redundant amendment is that the original text passage reads as "I had called upon my friend, Mr. Sherlock Holmes, one day in the autumn of last year, …".

5. Conclusion

We have presented a Web-based Japanese language learning environment, which offers engaging annotation exercises through a visually enhanced sentence representation. The current user interface design is the result of several iterative development cycles, which included feedback rounds with volunteer language students.

In the future we are going to continue to improve the user experience. For that purpose we intend to have our system soon ready for more widespread experimental use in language classrooms to obtain further feedback, which is also essential for issue tracking and system stabilization. Once we have reached the desired level of maturity, we plan to make the environment available on GitLab.

Apart from the application to other languages, a more ambitious and long-term research target will be the extension of our annotation to incorporate the document-level representation of UMR to be able to model intersentential dependencies.

We will also experiment with different user interaction modalities with varying degrees of automatic linguistic analysis and annotation. In addition, we will consider other application scenarios for additional target user groups. The quite unique setting of annotation tasks for language learning certainly requires additional skills including metalinguistic knowledge that have to be taught to the students. This restricts the applicability of our methodology to certain user groups like, for instance, university students. On the positive side, this also significantly widens the potential user base to students of translation studies, literature studies, linguistics, etc. For example, conducting psycholinguistic experiments represents a fascinating challenge for future work.

We are also very curious about the results of analyzing the constructicon data, which we will collect from the students. Future work in this subfield will address the research question of an optimal interaction with LLMs (Feng et al., 2023b) to create a neuro-symbolic AI system (Wan et al., 2024).

We see a strong potential of personal idiolectal constructicons to become a foundation for the next generation of AI to reach the desired faculties of conceptualization (Singer, 2021), generalization (Hupkes et al., 2023), reasoning (Arkoudas, 2023), and self-reflection (Whitten, 2023) on the long road to self-awareness (Chandha, 2021).

[29] https://www.aozora.gr.jp/

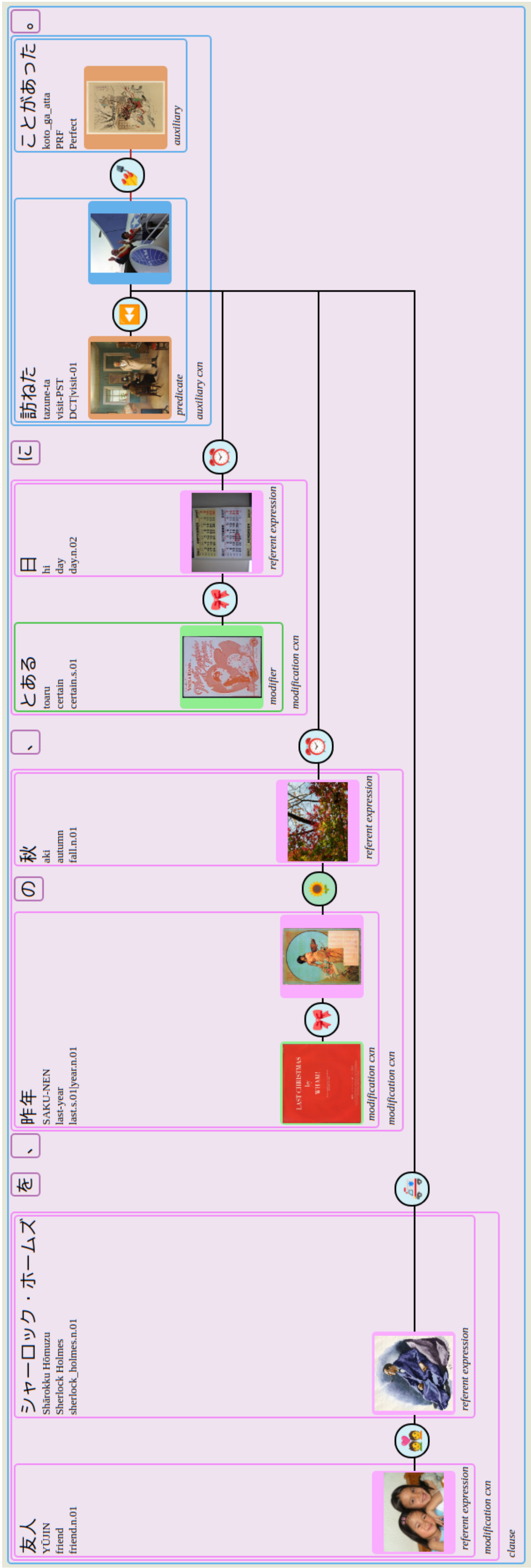


Figure 9: Complete annotation example.

6. Bibliographical References

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89, Vancouver, Canada.

Konstantine Arkoudas. 2023. GPT-4 can't reason. arXiv:2308.03762 [cs.CL].

A. E. Backhouse. 2016. *Using Japanese Synonyms*. Cambridge University Press.

Laura Banarescu et al. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online.

Arne Bewersdorff et al. 2024. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. arXiv:2401.00832 [cs.AI].

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

Francis Bond et al. 2009. Enhancing the Japanese WordNet. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 1–8, Suntec, Singapore.

Stefano Borgo et al., editors. 2019. *Ontology Makes Sense: Essays in Honor of Nicola Guarino*, volume 316 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Richard John Bowring and Haruko Uryu Laurie. 1992. *An Introduction to Modern Japanese*. Cambridge University Press, Cambridge, UK.

James Breen. 2004. JMdict: A Japanese-Multilingual dictionary. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 71–79.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020. Fatality killed the cat or: BabelPic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4680–4686, Online.

Saikiran Chandha. 2021. What if AI becomes self-aware. Express Computer.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press.

William Croft. 2002. *Typology and Universals*, 2nd edition. Cambridge Textbooks in Linguistics. Cambridge University Press.

William Croft. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge Textbooks in Linguistics. Cambridge University Press.

William Croft. 2023. Word classes in Radical Construction Grammar. In *The Oxford Handbook of Word Classes*. Oxford University Press.

William Croft and D. Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Jia Deng et al. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL.

Holger Diessel. 2023. *The Constructicon: Taxonomies and Networks*. Elements in Construction Grammar. Cambridge University Press.

Dagmar Divjak. 2019. *Frequency in Language: Memory, Attention and Learning*. Cambridge University Press.

Zhendong Dong and Qiang Dong. 2006. *HOWNET and the computation of meaning*. World Scientific.

Jonathan Dunn. 2022. Replication data for: Exposure and emergence in usage-based grammar: Computational experiments in 35 languages. DataverseNO.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database and some of its applications*. MIT Press.

Lydia Feng et al. 2023a. Widely interpretable semantic representation: Frameless meaning representation for broader applicability. arXiv:2309.06460 [cs.CL].

Zhangyin Feng et al. 2023b. Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications. arXiv:2311.05876 [cs.CL].

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16.3:235–250.

Michael Ginn et al. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada.

Jens E. L. Van Gysel et al. 2021. Designing a Uniform Meaning Representation for natural language processing. *KI – Künstliche Intelligenz*, 35:343–360.

Yoko Hasegawa. 2015. *Japanese: A Linguistic Introduction*. Cambridge University Press, Cambridge, UK.

Yoko Hasegawa, editor. 2018. *The Cambridge Handbook of Japanese Linguistics*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language*, 86:663–687.

Martin Hilpert. 2014. *Construction Grammar and its Application to English*. Edinburgh University Press.

Thomas Hoffmann and Graeme Trousdale, editors. 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press.

Dieuwke Hupkes et al. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174.

Shoichi Iwasaki. 2013. *Japanese*, revised edition. John Benjamins, Amsterdam, The Netherlands.

Stefan Kaiser et al. 2013. *Japanese: A Comprehensive Grammar*, 2nd edition. Routledge, London and New York.

Michiel Kamermans. 2010. *An Introduction to Japanese – Syntax, Grammar & Language*. SJGR Publishing, Rotterdam, The Netherlands.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain.

Yoko Kudo. 2011. Modified Hepburn romanization system in Japanese language cataloging: Where to look, what to follow. *Cataloging & Classification Quarterly*, 49(2):97–120.

Mathieu Lafourcade, Alain Joubert, and Nathalie Le Brun. 2015. *Games With a Purpose (GWAPSs)*, chapter The JeuxDeMots Project – GWAPs and Words. Wiley.

Young-Suk Lee et al. 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.

Jens Lehmann et al. 2015. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2024. Cross-modal retrieval for knowledge-based visual question answering. arXiv:2401.05736 [cs.CL].

Hugo Liu and Push Singh. 2004. ConceptNet: A practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Benjamin Lyngfelt, editor. 2018. *Constructicography: Constructicon development across languages*, volume 22 of *Constructional Approaches to Language*. John Benjamins.

Hiroshi Matsumoto. 2007. Peak learning experiences and language learning: A study of American learners of Japanese. *Language Culture and Curriculum - LANG CULT CURRIC*, 20:195–208.

Yoshiko Matsumoto. 1997. *Noun-Modifying Constructions in Japanese: A frame semantic approach*. John Benjamins, Amsterdam, The Netherlands.

George A. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Yoshiko Mori. 2014. Review of recent research on kanji processing, learning, and instruction. *Japanese Language and Literature*, 48(2):403–439.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

David R. Mortensen et al. 2023. Generalized glossing guidelines: An explicit, human- and machine-readable, item-and-process convention for morphological annotation. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 58–67, Toronto, Canada.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli et al. 2021. Ten years of BabelNet: A survey. In *Proceedings of IJCAI 2021*, pages 4559–4567.

Kyoko Ohara. 2014. Relating frames and constructions in Japanese FrameNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2474–2477, Reykjavik, Iceland.

Kyoko Ohara et al. 2003. The Japanese FrameNet project: A preliminary report. In *Proceedings of Pacific Association for Computational Linguistics (PACLING'03)*, pages 249–254.

Toshio Ohori. 2001. *Clause integration as grammaticalization: a case from Japanese Tokoro-complements*, pages 279–301. Kuroshio, Tokyo, Japan.

Anna Papafragou, John C. Trueswell, and Lila R. Gleitman, editors. 2022. *The Oxford Handbook of the Mental Lexicon*. Oxford University Press.

Simon Paxton. 2019. Kanji matters in a multilingual Japan. *The Journal of Rikkyo University Language Center*, 42:29–41.

Sameer Pradhan et al. 2022. PropBank comes of Age—Larger, smarter, and more diverse. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington.

James Pustejovsky et al. 2016. The development of multimodal lexical resources. In *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*, pages 41–47, Osaka, Japan.

Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. 1998. MindNet: Acquiring and structuring semantic information from text. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1098–1102, Montreal, Quebec, Canada.

Tanja Samardžić, Robert Schikowski, and Sabine Stoll. 2015. Automatic interlinear glossing as two-level sequence classification. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 68–72, Beijing, China.

Gadi Singer. 2021. Conceptualization as a basis for cognition – human and machine: A missing link to machine understanding and cognitive AI. Towards Data Science.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706.

Gilles Sérasset. 2015. DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF. *Semantic Web Journal*, 6(4):355–361.

Koichi Takezawa. 1993. Secondary predication and locative/goal phrases. In Nobuko Hasegawa, editor, *Japanese Syntax and Comparative Grammar*, pages 45–77. Kurosio, Tokyo, Japan.

Chiaki Taoka. 2000. *Aspect and argument structure in Japanese*. Ph.D. thesis, University of Manchester.

Tobias Ungerer and Stefan Hartmann. 2023. *Constructionist Approaches: Past, Present, Future*. Elements in Construction Grammar. Cambridge University Press.

Zishen Wan et al. 2024. Towards cognitive AI systems: a survey and prospective on neuro-symbolic AI. arXiv:2401.01040 [cs.AI].

Ryuichi Washio. 1997. Resultatives, compositionality and language variation. *Journal of East Asian Linguistics*, 6:1–49.

Allison Whitten. 2023. AI agents that "self-reflect" perform better in changing environments. Human-Centered Artificial Intelligence, Stanford University.

Jan Wielemaker, Zhisheng Huang, and Lourens Van Der Meij. 2008. SWI-Prolog and the Web. *Theory and Practice of Logic Programming*, 8(3):363–392.

Jan Wielemaker et al. 2012. SWI-Prolog. *Theory and Practice of Logic Programming*, 12(1-2):67–96.

Werner Winiwarter and Bartholomäus Wloka. 2022. VISCOSE – a kanji dictionary enriched with VISual, COmpositional, and SEmantic information. In *7th Workshop on Cognitive Aspects of the Lexicon (CogALex-VII)*.

Bartholomäus Wloka and Werner Winiwarter. 2021a. AAA4LLL – Acquisition, Annotation, Augmentation for Lively Language Learning. In *3rd Conf. on Language, Data and Knowledge (LDK2021)*.

Bartholomäus Wloka and Werner Winiwarter. 2021b. DARE – a comprehensive methodology for mastering kanji. In *23rd Intl. Conf. on Information Integration and Web Intelligence (iiWAS2021)*, pages 427–435.

Etsuyo Yuasa and Jerrold M. Sadock. 2002. Pseudo-subordination: a mismatch between syntax and semantics. *Journal of Linguistics*, 38:87–111.

Xingyuan Zhao et al. 2020. Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5397–5408, Barcelona, Spain (Online).

Michael Zock. 2022. The mental lexicon: A blueprint for the dictionaries of tomorrow? *Frontiers in Psychology and Artificial Intelligence*, 5.

Michael Zock and Chris Biemann. 2020. Comparison of different lexical resources with respect to the tip-of-the-tongue problem. *Journal of Cognitive Science*, 21(2):193–252.

Michael Zock, Paul Sabatier, and Line Jakubiec. 2008. Message composition based on concepts and goals. *International Journal of Speech Technology*, 11(3-4):181–193.

7. Language Resource References

Princeton University. 2012. *WordNet 3.1*. Princeton University, ISLRN 379-473-059-273-1.

Individual Text Corpora Predict Openness, Interests, Knowledge and Level of Education

Markus J. Hofmann[1], Markus T. Jansen[1], Christoph Wigbels[1], Benny B. Briesemeister[2], Arthur M. Jacobs[3]

[1]University of Wuppertal, [2]IU International University, [3]Free University Berlin
[1]Gaußstrasse 20, 42119 Wuppertal, Germany; [2]Juri-Gagarin-Ring 152, 99084 Erfurt, Germany;
[3]Habelschwerdter Allee, 14195 Berlin, Germany; {mhofmann, mjansen, christoph.wigbels}@uni-wuppertal.de, benny.briesemeister@iu.org, ajacobs@zedat.fu-berlin.de

Abstract

Here we examine whether the personality dimension of openness to experience can be predicted from the individual google search history. By web scraping, individual text corpora (ICs) were generated from 214 participants with a mean number of 5 million word tokens. We trained word2vec models and used the similarities of each IC to label words, which were derived from a lexical approach of personality. These IC-label-word similarities were utilized as predictive features in neural models. For training and validation, we relied on 179 participants and held out a test sample of 35 participants. A grid search with varying number of predictive features, hidden units and boost factor was performed. As model selection criterion, we used R^2 in the validation samples penalized by the absolute R^2 difference between training and validation. The selected neural model explained 35% of the openness variance in the test sample, while an ensemble model with the same architecture often provided slightly more stable predictions for intellectual interests, knowledge in humanities and level of education. Finally, a learning curve analysis suggested that around 500 training participants are required for generalizable predictions. We discuss ICs as a complement or replacement of survey-based psychodiagnostics.

Keywords: Big Five, PPIK theory, web tracking, predictive modeling. language models.

1. Introduction

While web tracking data are frequently used for individualized commercials and user profiling (Ermakova et al., 2018), they have not yet been used to predict diagnostic data from psychometric surveys. Here we rely on the google search history to predict openness to experience from a Big Five survey. Our basic hypothesis is “you are what you read” (cf. Schaumlöffel et al., 2018), which we test by the similarity of the googled homepages to label words defining personality.

We collected a total of 214 google search histories from 214 participants and used web crawling to generate individual text corpora (ICs, Hofmann et al., 2020). We held out a test sample of 35 participants and used 179 for training and validation (Figure 1A).

The semantic structure of each participant’s reading material was defined by a word2vec model (Figure 1B; Mikolov et al., 2013), which is a relatively simple neural language model. In skip-gram mode, hidden units are trained to predict the surrounding words by each present word. After training, each word obtains a vector representation that defines its meaning by the language contexts, in which it is typically embedded. To compute the semantic similarity of two words, each entry in this vector is then considered as a dimension in a multidimensional space and the cosine between the two vectors is taken to define semantic similarity.

For defining personality, we relied on the lexical approach, which goes back to Sir Francis Galton: “the most important individual differences in human transactions will come to be encoded as single terms in (…) language(s)” (quoted from Goldberg, 1993, p. 26). Therefore, we started with adjectives that were taken to construct Big Five surveys (Ostendorf, 1990) and expanded them by similar verbs and nouns as word labels (cf. Westbury et al., 2015). A similar approach has been successfully applied, for example, to predict the Big Five of fictive characters such as Harry Potter or Voldemort (Jacobs, 2019, 2023). We computed the cosine similarity of the 2500 most frequent words of each IC to these word labels (Figure 1C). Then we averaged across these 2500 words to obtain the similarity of the label words to the individual reading materials of each participant (Figure 1D).

These IC-label similarities were then used as predictive features for between-subject neural models predicting the survey-based openness to experience from the Big Five surveys (Figure 1E).

We performed a grid search with 30 to 100 label words and a wide variety of model complexity of the neural models. For model selection, we use the explained variance in the validation set penalized by the absolute difference between the training and validation set. Then we examined the predictive performance of the selected model and an ensemble model with the same architecture in the test set. We also tested some predictions from Ackerman's (1996) theory of intellectual development. He proposes that the personality feature of openness to experience often leads to the development of intellectual interests. It should also foster knowledge in the humanities, which is often apparent in individuals open to any type of experience. We also compared the neural-model-based openness with survey-based openness for the prediction of level of education. Finally, we performed a learning curve analysis to estimate the required sample size of this ongoing data collection.

The Workshop on Cognitive Aspects of the Lexicon, pages 14–25
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

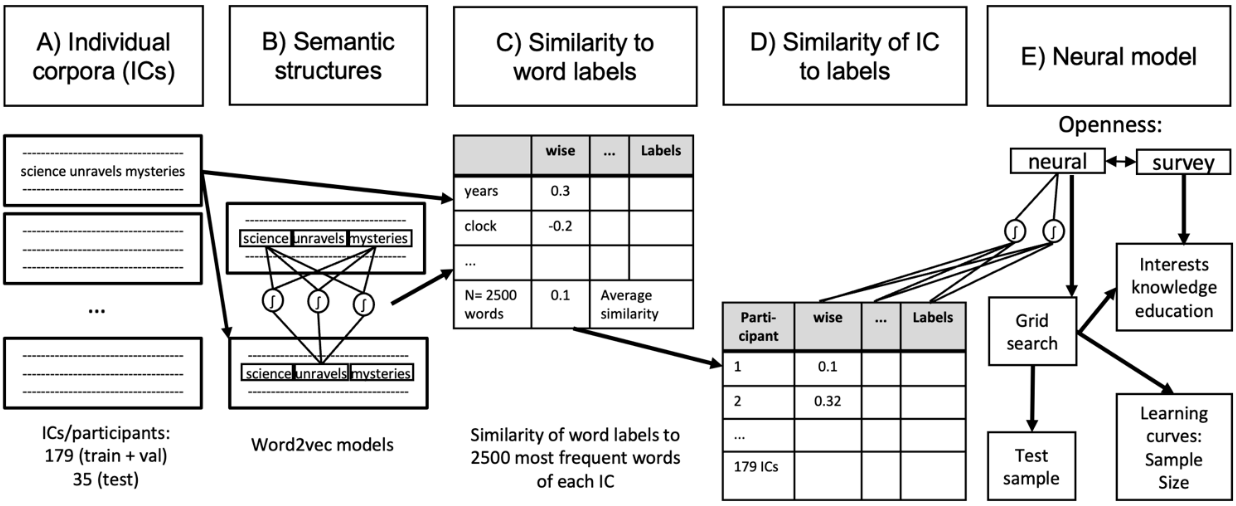


Figure 1: Overview of the present study (see Introduction).

2. Related and present work

The Big Five factors of personality have been associated with a vast number of psychological traits (McCrae and Costa, 1987). The OCEAN model characterizes subjects on the five personality dimensions of *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness* and *Neuroticism*. *Openness* characterizes subjects as inventive, curious and with broad aesthetic interests – subjects open to experience expose themselves to diverse environments (rather than following routines) and they are attentive to their own and other emotions. *Conscientiousness* can be subsumed as orderliness and self-discipline, for instance, and thus it predicts academic achievement and job performance (Hurtz & Donovan, 2000). *Extraversion* characterizes subjects as enthusiastic, energetic and adventurous. Together with conscientiousness both favor psychological well-being (Anglim et al., 2020). *Agreeableness* is important for establishing and sustaining friendships and other kinds of relationships (Harris & Vazire, 2016). *Neuroticism* describes participants that are nervous, insecure, frequently complaining and stress-sensitive. The predominantly negative emotions thus promote the development of affective disorders, while the lack of neuroticism is considered as emotional stability (Lyon et al., 2021).

Pennebaker and King (1999) set initial benchmarks for the prediction of personality relying on diary studies (McCrae and Costa, 1987). For instance, the frequency of using words with more than six letters provided the largest correlation with openness to experience (r = .16, Pennebaker and King, 1999). While Yarkoni (2010) later confirmed such findings using internet blogs, Schwartz et al. (2013) showed that only some of these findings are reproducible using Facebook posts – they suggested that the sparse sample of simple word counts may be the reason for the relatively poor amount of explained variance. Schwartz et al. (2013) showed that only some of these findings are reproducible using Facebook posts – they suggested that the sparse sample of simple word counts may be the reason for the relatively poor amount of explained variance. Schwartz et al. (2013) compared this approach with a topics modeling approach that provides one semantic structure for the whole sample of participants. They computed the topics that participants with particular personality features frequently use and showed that participants with high openness use topics that contain words such as “writing”, “read” and “poem”. Their topics models provided larger correlations with the survey-based openness (r = 0.38) than Pennebaker and King's (1999) seminal work (Eichstaedt et al., 2021). A similar correlation (r = .43) was reached by Kosinski et al. (2013), who analyzed Facebook likes by an LSA-based approach. While Schwartz et al. (2013) discussed that an upper limit of reproducible Pearson correlations could range between .3 and .4, Azucar et al. (2018) later performed a meta-analysis based on 14 studies and showed that the meta-analytic correlation of social media data with openness is r = .39.

Eichstaedt et al. (2021) called Pennebaker and King's (1999) work a closed vocabulary approach of only theory-based word labels. In contrast, they used an open vocabulary approach exploring all available word types. In the present study, we use a relatively closed vocabulary, which is based on Big Five adjectives form a lexical approach to personality (Goldberg, 1993; Ostendorf, 1990). Moreover, we expand these adjectives by label words from other word classes. To find verbs and nouns for the personality-descriptive adjectives that are frequently co-occurring with the adjectives in selected syntactic dependencies, we rely on JoBimText using the German parsed lemma database (Biemann & Riedl, 2013)[1].

[1] See e.g., http://ltmaggie.informatik.uni-hamburg.de/jobimviz/for a web demonstration.

Here we aim to build a theory-based, relatively closed vocabulary approach, because we test several hypotheses derived from Ackerman's (1996) theory, who proposed that intellectual development is a *Process*, in which *Personality* creates specific *Interests* and crystallized *Knowledge* (PPIK theory, Ackerman, 1996). He relies on Holland's (1959, p. 36) theory proposing that humans with a high intellectual interest have "marked needs to organize and understand the world". Based on this large theoretical framework, Ackerman (1996) was able to explain that openness provides a correlation with intellectual interests (cf. Kandler and Piepenburg, 2020). Similarly, Rolfhus and Ackerman (1996) showed that openness to experience is particularly correlated with knowledge in the humanities (e.g., literature, philosophy, Schipolowski et al., 2013). It was assumed that specialized knowledge structures emerge from fluid intelligence. This general reasoning ability further requires the investment of time into a particular field of knowledge (Cattell, 1987; von Stumm & Ackerman, 2013). Therefore, more specific knowledge develops over time and should diversify over the life span (Jacobs & Kinder, 2022; Watrin et al., 2021). During this development, openness on the one hand influences crystallized intelligence, but can on the other hand also foster the development of fluid abilities (Ziegler et al., 2012). Though this theory provided a perspective on intellectual development over time, it has rarely been tested successfully in psychology, because longitudinal data are often missing (but cf. Ziegler et al., 2018).

ICs may be a useful approach to test this theory, because the google search histories contain time stamps and on average our participants started to google in 2015. While the present study starts by estimating openness as a relatively stable trait, once we established a functional predictive model, we plan to examine intellectual development in a longitudinal perspective.

While the previous computational approaches to personality relied on one language model for the whole sample of participants (e.g., Eichstaedt et al., 2021), Hofmann et al. (2020) proposed that an individual corpus, from which a predictive language model is trained, reflects a sample of individual human experiences. They used ICs generated from the reading of two participants on a tablet for two months to train word2vec models reflecting the individual long-term memory systems of these participants. They compared the ICs of the two participants to a standard corpus for predicting reading performance in an eye-tracking study. Though the ICs were comparably small with 300/500K word tokens, only the ICs were able to successfully predict fast memory retrieval during reading in this rather limited data set.

In recent years, predictive modeling has found its way into psychology. For instance, Koutsouleris et al. (2016) relied on survey and other data to predict treatment outcomes. While such classifier approaches are frequently used, there are also a few regression approaches on continuous data (e.g., Jankowsky et al., 2023). We started with the explained variance in the validation set. Overfitting is given when more variance is explained in the training than in the validation set, while the reverse is true for underfitting. Therefore, we used the explained variance in the validation set as model selection criterion penalized by the absolute difference between training and validation R^2's to penalize over- and underfitting.

3. Methods

3.1 Participants and surveys

At the time point of this ongoing data collection at which we report the analyses, a total of 295 people participated in the study. We excluded 81 participants who did not provide an appropriate Google search history file or less than 2500 word types in their ICs. The final set of 214 participants were adult German native speakers without any language disorders (e.g., dyslexia) who actively used a Google account for at least one year (149 females; age: M ~ 28; SD ~ 8[2]). Subjects either received course credits or 10€ for participation[3].

The 60-90 min online survey started with demographic data, where participants reported age, gender and level of education (1 = no academic degree, 2 = secondary modern school [Hauptschule], 3 = intermediate school [mittlere Reife], 4 = technical college entrance qualification [Fachabitur], 5 = general university entrance qualification [Abitur], 6 = academic degree). They were instructed to browse to https://takeout.google.com, log into their Google account, and download a myactivity.json file, which was later uploaded on the survey homepage.

We relied on psychological surveys available under Creative Commons Licenses to facilitate later re-use. For personality assessment, we used the mean ratings of the Big Five Aspect Scales (BFAS-G, Mussel and Palaecke, 1996), which consist of 100 statements such as "I have fun enjoying nature" for openness. Participants rated whether this statement applies on a 7-point Likert scale. Item consistency was acceptable (Cronbach's α: O = .79; C = .83; E = .87; A = .85; N = .92).

General and domain-specific knowledge was based on the BEFKI GC-K (Schipolowski et al., 2013). In addition to the original 12 questions on the knowledge areas of natural science (biology, geography medicine, physics), humanities (art, literature, philosophy, religion) and social science (finance, history, law, politics), we created two additional questions for each knowledge domain. Four multiple choice answers were available for a total of 36

[2] To facilitate anonymization, participants reported approximate age ranges.

[3] This research was and will be funded by the German Research Foundation (HO 5139/4-2 and 6-1).

questions and the number of correct answers (per knowledge area) were examined. As is usual for such short scales, item consistencies were in part questionable for the knowledge areas (humanities = .56; social science = .50; natural science = .76), while the overall scale representing general knowledge was acceptable (α = .76).

We additionally addressed crystallized intelligence by a general intelligence screening (AIT Satow, 2017; Cronbach's α = .84). In each of the 18 items, a list of three words is presented and a fitting fourth word has to be selected from a list of five options (e.g., here - then - maybe: warm, big, now, nice and run). A screening of fluid intelligence was assessed by the syllogism task of this test. The 15 items consist of two premises, e.g., no rectangle is a circle; all squares are rectangles. Participants have to infer on one of four options: no square is a circle; all squares are quadrilaterals; no square is a quadrilateral; some quadrilaterals are rectangles. We here also obtained good internal consistencies (α = .85). We used the sums of correct answers for the prediction by openness to experience.

Leisure interests were examined by mean ratings on the 5-point Likert scale of the FIFI-K (Nikstat et al., 2018). It consists of 67 questions concerning everyday activities. The second-level factor of intellectual interests consists of 10 short statements such as "Watching news/reading newspaper" and participants answer how frequently (F, α = .64) they perform the activity and how much they like it (L, α = .62).

3.2 Language modeling

We used python3 for language modeling. The myactivity.json files were constrained to text information, anonymized, tokenized, filtered to obtain the German individual corpora and stemmed. The ICs provided a mean token number of 5,028,586 (SD = 7,961,353).

We excluded stopwords and HTML codes and used Genism 3 to train skip-gram models with 300 hidden units for each IC/participant (window size = 2, training epochs = 10, minimum word frequency = 3, Hofmann et al., 2020; Rehurek & Sojka, 2010). We extracted the 2500 most frequent words of each IC and computed the cosine similarity to each label word. Then we computed an average similarity across the 2500 words to each label word to obtain the similarity of each IC to each label.

Label word selection started with a pool of 430 personality-descriptive adjectives (Ostendorf, 1990, pp. 168-177). The basic idea of this lexical approach is that the description of personality is reflected in language (Goldberg, 1993). Following this approach, Ostendorf (1990) performed an extensive set of factor analyses to generate this Big Five word list. We assigned each adjective label one or more Big Five personality dimensions based on the factor loadings on the respective dimension. For feature expansion, we relied on JoBimText (Biemann and Riedl, 2013) for finding verbs and nouns, which are frequently co-occurring with the adjectives in specific syntactic dependencies. We selected these syntactic structures manually, such that the verbs and nouns were intuitively similar to the selected adjectives. We assigned the verbs and nouns the personality dimension, they were derived from, and excluded labels that occurred in less than 50% of the ICs, leaving us with a total of 398 label words for openness to experience. Note that we computed the similarity within the individual semantic structure that were delivered by the word2vec models. Therefore, when the respective label word was not contained in the IC, the IC-to-label-word similarity was defined as zero. In other words, here we assume orthogonality of this IC to the respective word label.

The average similarities across the 2500 most frequent words of each IC to each label word were submitted to JMP Pro 17 for predictive modeling, where we used the similarity of the ICs to the label words as predictors.

3.3 Predictive modeling

To examine whether ICs can predict openness to experience, we first built a stratified training, validation and test sample consisting of 143, 36 and 35 of the participants, respectively. For assuring that the predictive features provide a similar variability in the training, validation and the test sample, we performed a k-means cluster analysis on the 398 word IC-label similarities, setting the cluster size to 3 (e.g., Burden et al., 2000). We stratified the samples for cluster membership and distance.

The similarity of each IC to the label words were then used as predive features. For feature selection, we performed a random-forest analysis based on 10^{30} trees predicting openness by the 398 label words assigned to openness. We ranked the label word similarities by the proportion of trees, they were contained in, and examined the 30 to 100 highest-rank label words in the grid search (step size 10). Unless otherwise noted, all random seeds of the present study were set to 1.

The second hyperparameter in the grid search was neural model complexity: We built neural models with varying numbers of hidden units (1 to 10, step size: 1; 20-100, step size 10; 150 to 500 step size 50) and boost factors of 0 (no boosting) to 5. During boosting, another neural model is fitted to the residuals of the preceding model. The model complexity variable starts with the simplest model with 1 TanH unit without boosting (complexity = 1), followed by the boosted variants (e.g., boost = 1 corresponds to a complexity of 2). The maximum model complexity is 162 and corresponds to 500 hidden units with a boost of 5 (see Figure 2). We repeated all models 10 times with different random seeds, but the starting random seed of each hyperparameter set was kept constant at 1. Thus 1,620 neural models were fitted for each number of predictive word labels. With the 30 to 100 predictive features, this resulted in a total of 12,960 neural models for the grid search. For this search, we pooled the training and validation set and used 5-fold cross validation.

We propose a new model selection strategy that is based on a model evaluation criterion, which we call *Mis-Fit Penalized R^2*:

$$MFPR^2 = R^2_{Val} - (| R^2_{Train} - R^2_{Val} |) \qquad [1]$$

The procedure for model selection was the following:

1. Calculate generalized mean R^2_{Train} and R^2_{Val} values over the 10 neural models with different random seeds for each hyperparameter set from the grid search.
2. Compute a spline function for each number of label words, which compares model complexity on the x axis to $MFPR^2$ on the y axis.
3. Select the hyperparameters that are most closely to the maximum of the $MFPR^2$ spline (cf. x axis, Figure 2) and that provides the highest $MFPR^2$ (y axis on Figure 2).

The selected hyperparameters will be further evaluated by additional 100 neural models fitted with randomly chosen seeds, from which we also compute the average probabilities to obtain a more stable ensemble model.

For the learning curve analysis, we used the training, validation and test sets, as initially split by the cluster-based stratification. We kept the validation and training sets constant with 36 and 35 participants, respectively, and started our evaluation with 35 training rows. To obtain a relatively homogenous sequence of training cases, the training set was split into 10 subsets, stratified for cluster membership, distance and openness. These subsets were presented sequentially (cf. Ouyang et al., 2021). For each number of training rows, we fitted 100 neural models with different random seeds.

To fit the mean average error (MAE) of the training and test set as a function of the training rows (TR), we fit a power function with an intercept (Viering and Loog, 2021; starting values: a = .5, b = -.5; c = -1).

$$MAE = a * TR^{-b} + c \qquad [2]$$

While model selection was based on the generalized R^2 values computed from the training and validation sets (Figure 2), for model evaluation we additionally report R^2 values computed from the Pearson correlation coefficient of the samples to facilitate comparability with previous studies predicting personality.

4. Results

4.1 Model selection and evaluation

When examining the results of the grid search in Figure 2, local spline maxima of $MFPR^2$ were found at medium model complexity. Spline functions suggested that 60 predictors provide the largest $MFPR^2$ values, while the second-best $MFPR^2$ splines were obtained with 50 label words. Based on these spline functions, we selected the hyperparameter set providing the highest $MFPR^2$ values (y axis) that provide the lowest distance to the spline maximum (x axis in Figure 2). When $MFPR^2$ computed from the training and validation samples generalizes to the test sample, we thus expect better results for the best as compared to the second-best model.

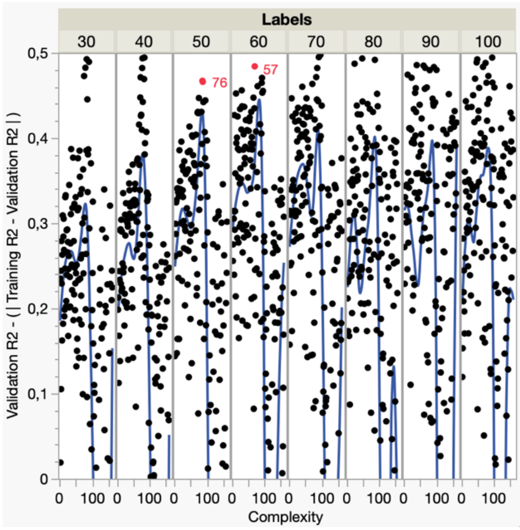


Figure 2: Results of the grid search. We used 30-100 IC-label word similarities as predictive features and examined model complexity on the x axis (1-500 hidden units, boost 0-5). On the y-axis, we display $MFPR^2$ as the model selection criterion.

The best model with 60 labels provided 30 hidden units and a boost of 3 (model 57 in Figure 2). For a closer model evaluation, we then fitted 100 models with randomly varying seeds. With 5-fold cross validation, we obtained a mean r = .28 between predicted and observed openness in the test set (SD = .11). 32 of the 100 correlations were significant ($P \leq .05$). We did not apply Bonferroni correction, because we were interested in the generalizability of the predictions of this hyperparameter set, rather than interpreting single significant correlations. There were only 35 participants in the test sample, thus the significance threshold is .35 (Roberts et al., 2007, p. 314, cf. below) and the expected number of significant correlations is 5. Therefore, we consider this an appropriate number of significant tests demonstrating generalizability of the selected hyperparameter set. To examine whether the results change with greater training and smaller validation samples, we also performed 10-fold cross validation. It revealed a mean r = .25 (SD = .15). 27 of the 100 correlations were significant.

The second-best model was trained by 50 label words and contained 60 hidden units without boosting (model 76 in Figure 2). We also examined this model to probe our model selection strategy. With 5-fold cross-validation, we obtained a mean correlation of the predicted with the empirical openness of r = .21 in the test set (SD = .13). In this sample, 18 of the 100 correlations were significant. In the 10-fold cross-validation, we also observed a mean r = .21 (SD = .15). 24 of the 100 correlations were significant.

In sum, the best model, which was selected from the training and validation sample, provided a reasonable fit in the test sample. It performed better than the second-best model, which demonstrates the generalizability of the $MFPR^2$ model selection strategy to the test sample.

For further model testing, we selected the best neural model from the grid search (random seed = 1630049, 5-fold). When examining correlations between the predicted and observed values for the pooled training and validation samples, and the test sample alone, this model provided a moderate amount of overfitting, as examined by Pearson-based explained variances ($R^2_{Train+Val}$ = .50; R^2_{Test} = .35).

To obtain more stable predictions, we also built an ensemble model by averaging the predictions over the 100 neural models. As this ensemble model contains also models with a poorer fit, this model provided more overfitting ($R^2_{Train+Val}$ = .62; R^2_{Test} = .20).

4.2 Psychometric examination

When examining the overall sample, the ensemble model and the selected best model provided high correlations of neural-model-based with survey-based openness (bold in Table 1). These correlations were higher than correlations of self- and peer-reported personality scores (cf. McCrae and Costa, 1987, Table 6). In psychometric terms, we can conclude that the convergent validity of the ensemble neural model's openness is higher than the inter-rater agreement of other studies. Moreover, the neural modeling openness scores provided lower correlations with the other Big Five dimensions than the survey-based openness (bold in Table 1), except from agreeableness in the test sample, which tended to provide a (non-significant) negative correlation with neural openness. For all other cases, the neural openness scores showed larger divergent validity than the survey-based openness data. In sum, the neural modeling openness revealed better convergent and divergent validity than the survey-based openness.

When examining intellectual interests, larger correlations were obtained for the reported liking than for the frequency of doing intellectual leisure activities (italic in Table 1). Also larger correlations were obtained for the ensemble neural model predicting the liking of intellectual leisure activities. For the overall sample, survey-based openness provided higher correlations with the liking of intellectual activities than ensemble-based openness, but for the test sample, higher correlations were obtained for both neural models as compared to the survey-based openness.

As also predicted by the PPIK theory (Rolfhus and Ackerman, 1996, Table 5), we observed significant correlations of all openness scores with the knowledge in humanities in the overall sample (italic in Table 1). These openness correlations were higher than the correlations with knowledge in natural and social science (data not shown). Only the survey-based openness reached a significant correlation with social sciences (r = .18).

For general knowledge, which reflects crystallized intelligence, only the survey-based openness provided a significant correlation with the sum of all correctly answered knowledge questions. Also, for the other intelligence tasks (not shown in table), only survey-based openness provided a significant correlation with crystallized and fluid intelligence (both r's = .18).

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Ensemble neural (O)		.86	**.75**	**.21**	**.24**	**.14**	**-.15**	.22	***.41***	***.25***	.12	***.17***
2. Selected neural (O)	.84		**.68**	**.19**	**.21**	**.15**	**-.14**	.16	***.33***	***.21***	.09	***.15***
3. Openness (O, survey)	**.45**	**.59**		**.25**	**.32**	**.29**	**-.20**	.34	***.46***	***.31***	***.26***	***.17***
4. Conscientiousness	**.18**	**.25**	.30		.30	-.00	-.38	.06	.05	.02	.09	.00
5. Extraversion	**.01**	**.12**	.37	.23		.06	-.32	-.02	-.05	.05	-.03	.02
6. Agreableness	**-.27**	**-.27**	.00	.14	.12		.10	-.03	.05	-.04	-.02	.02
7. Neuroticism	**-.09**	**-.08**	-.30	-.28	-.09	-.02		-.04	-.07	-.17	-.23	-.09
8. Intellectual Interest (F)	.03	.17	.25	.30	-.08	.03	-.05		.61	.18	.15	-.02
9. Intellectual Interest (L)	***.41***	***.37***	***.35***	.06	-.18	-.16	.10	.16		.27	.21	.19
10. Knowledge humanities	-.01	-.03	.17	-.27	.08	-.10	-.25	-.15	.03		.81	.34
11. General knowledge	-.06	-.02	.07	-.03	.08	-.02	-.48	-.05	-.13	.77		.35
12. Level of education	.14	.00	-.00	-.00	.10	.03	-.01	-.33	.38	.13	.10	

Table 1: Below diagonal correlations for test sample (N = 35), above diagonal complete sample (N = 214). For the overall (test) sample, correlations crossing an r = .14 (.35) are significant (P ≤ .05). Convergent and divergent validity scores (theoretical predictions) are printed in bold (italic).

The correlations were usually higher for the survey-based than for the neural-model-based openness scores, except for the liking of leisure activities in the test set. The level of education was predicted equally well with survey-based and ensemble-neural-model-based openness to experience.

4.3 Learning curves and sample size

To estimate a sample size at which full generalization should occur, we performed a learning curve analysis on the mean absolute error (MAE). For this analysis, we unpooled the training and validation samples, and report training and test performance. The ensemble model indicated large overfitting in terms of generalized R^2 scores (R^2_{Train} = .63; R^2_{Val} = .64 ; R^2_{Test} = .20), while the best model only provided a moderate amount of overfitting (R^2_{Train} = .49 ; R^2_{Val} = .57; R^2_{Test} = .35). Therefore, the best model was selected for the learning curve analysis.

Figure 3 displays the MAE values of 100 neural models fitted with randomly chosen seeds, which were fitted for each number of training rows (light colors). The dark lines indicate power functions fitted to the training and test data (Viering and Loog, 2021). With only a few training rows, the error is quite low in the training sample. Imagine, for instance, that a line fitted through two points will always fit perfectly (see e.g., Mehta et al., 2019). As in this case a lot of error variance is fitted, the corresponding errors in the test sample are high. With an increasing number of training cases, however, the learning curves for the training and test set approach each other. They should reach a common asymptotic value, when there is no generalization gap, which would indicate that the training cannot be fully generalized to the test data (e.g., Chao, 2011, Fig. 16). We observed crossed learning curves, which is a frequently observed phenomenon (Viering and Loog, 2021). We think that this is presently due to the limited number of training data rows of a training sample of up to 143 participants, which prevents an excellent fit for a larger number of training data rows. The training and test curves crossed between 500 and 600 participants. Therefore, we conclude that a training sample of around 500 participants should be sufficient to obtain fully generalizable results.

5. Discussion

5.1 Model selection and evaluation

We relied on label words to define personality and computed the similarity of each IC to these label words. By examining 30-100 label words and a wide variety of model complexity in a grid search, we were able to find a local maximum of our model selection criterion. We used the explained variance in the validation set penalized by the absolute difference in explained variance between the training and the validation samples. The latter term likewise penalizes over- and underfitting using the training and validation sample. Therefore, we hypothesized that it is an indicator of generalizable predictive performance.

To find areas in hyperparameter space that generally allow for good predictions, we computed spline functions over the average model selection criterion of 10 fitted models per hyperparameter set. We feel that such spline functions are a computationally relatively cheap way to identify an area in hyperparameter space that provides good model performance, as opposed to fitting more models per hyperparameter set. Otherwise, we would feel that this is necessary, because model performance varies considerably with the selected random seed (see section 4.1). With spline functions smoothed over many similar hyperparameter sets, however, an appropriate hyperparameter set can be based on many observations.

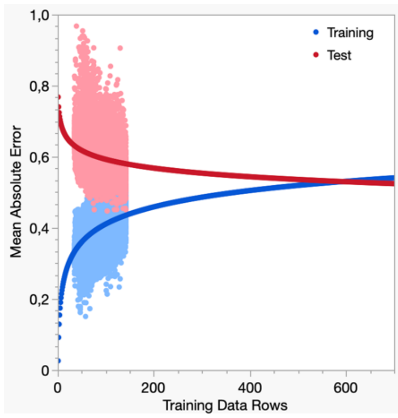


Figure 3: Learning curves. Light colors indicate MAE scores for 100 models fitted at each number of training data. Dark colors indicate fitted power functions.

The hypothesis that our model selection criterion, based on the training and validation set, also generalizes to the test set was confirmed. While also the second-best model found with this selection criterion was able to account for variance in the test sample, the best model clearly also performed better in the test sample, where it accounts for 35% of the variance. In addition to the best model, we also built an ensemble model with the same architecture by averaging prediction probabilities over 100 models with the same architecture, but randomly chosen seeds. Therefore, we demonstrate the generalizability of this hyperparameter set.

In sum, our AI system explains human personality as a learning system, in which the individual human experience, as sampled from ICs, is used to train a model of the semantic long-term memory system of each individual (Hofmann et al., 2020, 2022). Our basic hypothesis was that participants expose themselves to materials that reflect their personality (Schaumlöffel et al., 2018) and thus computed the similarity of theoretically well-founded personality-descriptive terms to the searched pages as a sample of individual human experience. Word2vec models provide intuitively valid similarities and can well explain human association ratings (e.g., Hofmann et al., 2018). Their simple embeddings are computationally well-defined and as opposed to large language models, they represent an epistemically translucent approach with greater explanatory value, though they are not fully deterministic (Hofmann & Jacobs, 2014). The training of these language models can be considered to reflect memory consolidation. We propose the word label similarities to the ICs as items of a psychological test to which the neural model of each participant responds. The between-subjects neural model then evaluates the output from the within-subject language models and can be considered as a psychodiagnostics model. And when considering the first neural model as a within-subjects layer and the second neural model as between-subjects layer, the output of the first layer can be symbolically interpreted – therefore, we also consider our approach as an explainable deep neural network model. In a nutshell, we consider our approach as a well-explainable AI in every part of our system, which is conceptually, methodologically and theoretically well-founded in psychological, language and predictive modeling approaches.

5.2 Psychometrics and sample size

Except for the prediction of survey-based openness in the test set, the ensemble model provided higher correlations with the survey data. We feel that this is a sound demonstration that the found hyperparameter set provides a solid approach to these data. Particularly with this ensemble model we were able to confirm the predictions of the PPIK theory that high openness to experience comes along with greater intellectual interest and more knowledge in humanities in the multiple choice knowledge test (Schipolowski et al., 2013). For crystallized intelligence, this prediction was not confirmed, much as for the prediction of fluid intelligence (Satow, 2017). Knowledge in humanities, however, can be predicted by neural-model-based openness – we think that this comes from an overlap between labels defining openness and knowledge in the humanities (e.g., wise, see Figure 4 below). We already started to experiment with predicting knowledge directly, and our preliminary evidence suggests that knowledge can be even better predicted by ICs than personality. It is interesting that we were not able to predict fluid intelligence (Satow, 2017). The text materials, the participants are searching, may reflect personality, but what they learn from the text might be better predicted by fluid intelligence. Therefore, future studies may use fluid intelligence as a covariate in order to address what participants learn from the texts. Fluid intelligence might also be worthwhile to be directly predicted, because we feel that even some of the largest language models have problems with reasoning, generalization and inference.

When examining the correlations between the neural-model predicted and the survey-based openness, we observed higher correlations than McCrae and Costa (1987) found between different raters. This is a well-known result pattern for language-model-based approaches to personality – but as opposed to Youyou et al (2015) we were able to show similar correlations in a hold-out test sample that was not used to fit the data. In any way, this shows good evidence of convergent validity. Also, we present sound evidence for divergent validity of the present approach, i.e. the correlations of neural-model-based openness with the other personality dimensions were usually lower than for the survey-based openness. The neural-model-based openness scores can thus be better separated from the other personality dimensions.

While already previous computational studies started to stretch the "correlational upper limit" of .3 in predicting behavior by personality (Roberts et al., 2007, p. 314), the present study outperforms all previous studies reviewed by Azucar et al. (2018). As opposed to these studies, however, the present study even comes to larger correlations for the independent test set that has not been used for model training.

Finally, we performed a learning curve analysis by examining the training and test error as a function of the number of training data rows. This suggests that ~500 participants may be sufficient for generalizable predictions. We hope that the presently apparent differences between the selected and the ensemble model will vanish with such a sample size.

We are quite optimistic that in the near future, there is less need for time-consuming diagnostic assessments. When for instance, a future therapist wants to get a quick idea of the psychic properties of the client, analyzing the search history might be sufficient to get a great deal of psychometric information. Therefore, we think that such neural models may complement diagnostic information from surveys or even replace it. Moreover, this rapidly obtained information may be used to form hypotheses, which can be tested more thoroughly by the classic diagnostic assessment.

When considering the IC-label-word similarities as diagnostic items, we obtained a good internal consistency for the 60 labels of the selected model (Cronbach's α = .89), which we consider a sound basis for diagnostic approaches relying on the internet search history.

Last but not least, surveys are self-reported explicit answers measuring personality. It is well-known, for instance, that some answers can be affected by factors such as socially desirable responding. Therefore, at least for some behavioral phenomena, so-called implicit approaches to psychological traits may be more predictive. For motivation psychology, for instance, projective testing is often favored as an implicit measure of psychological traits (Winter, 2010). In projective testing, a quite ambiguous picture is shown to the participants and they are required to write a short story about the picture, assuming that participants project their own traits onto the picture. For instance, when they use many achievement-related terms, participants are assumed to have a large achievement motive, which in turn predicts business success (Winter, 2010). We propose that openness computed from a sample of individual human experience also reflects the implicit semantic structure of the participants. Such a highly reproducible computational approach to implicitly defined personality may help to overcome the limitations of projective testing, such as low evaluation objectivity, i.e. the difference between different evaluators, though this disadvantage of projective testing has already been tackled by language modeling (Johannßen et al., 2019). With a larger sample, we hope that implicit neural modeling openness can predict other behavior better than with explicit survey-based openness. Examples of this could be the test sample's correlation of the ensemble model with the liking of intellectual activities, or in another previous study we showed that level of education is better predicted by implicit neural modeling openness to experience (Hofmann et al., 2023). Therefore, the aim of this line of research is not to provide perfect correlations with survey-based openness, but to define implicit psychological traits that may predict other external validation criteria better than explicit, survey-based predictors.

5.3 Nonlinear activation functions

To demonstrate the theoretical benefit of language-model-based neural models of personality, we also examined the variable importance of the selected neural model to select exemplary nonlinear functions that elucidate the inner workings of the selected neural model. For examining variable importance of single label words, we assumed that the input variables would be independent and assessed the change in predictive performance when the single label words are dropped from the selected neural model. Then we examined the face validity of the most important predictive features.

The label word "wise" (weise) was a very important predictor. When the similarity of the IC to this label word increased, we observed an approximately linear increase of openness (Figure 4). Thus, despite this actually nonlinear approach, such linear relations demonstrate the face valid interpretability of such a neural model.

The label "show" (zeigen) also provided a high variable importance. It was derived from an adjective providing high factor loadings on openness and extraversion. There was a relatively linear decrease in similarity with high openness. This was a bit unexpected, because participants open to experience could be assumed not only to be open to experience by themselves, but also would tend to show new experiences to others. As the influence of openness must be differentiated from extraversion, however, the linear decrease makes sense, because extraversion may be the critical personality dimension leading to the motivation to show things to others.

For the label "withdraw" (entziehen) there was a more nonlinear, negative influence. When this label is less similar to the IC, participants are more open to experience, but when the IC is more similar to this word than .1 participants tend to be less open to experience. This suggests that participants being more open rarely withdraw from an occasion providing new experiences.

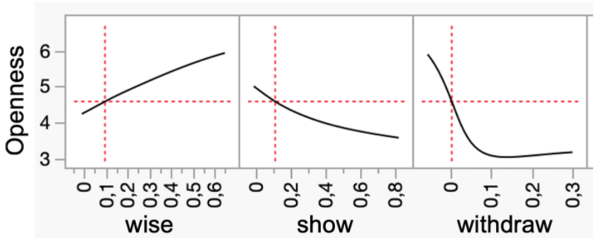


Figure 4: IC-label word similarities predicting openness.

5.4 Limitations and Outlook

The most obvious limitation of the present study is the limited number of participants, thus generalization is still limited. Therefore, we computed learning curves and estimate that the predictions should become stable when a sample of 500 participants is available. Nested cross validation may be used, which allows to rely on the whole sample. We are going to perform the same set of analyses with the other Big Five factors. The intelligence screening should also be complemented by a full assessment of intelligence Moreover, we also plan to predict area-specific knowledge with the present data. As the Google search history provides time stamps, a longitudinal perspective on intellectual and personality development will be possible.

We think that it is necessary to slightly change the stratification strategy, also stratifying for other dependent variables – in fact, lower explained variance may have been obtained for level of education, because it slightly varied between the training (M = 4.79, SD = .95), validation (M = 4.72, SD = .85) and test sample (M = 5.00, SD = .80). Moreover, greater representativeness (in lower levels of education) is desirable (for this self-reported external validation criterion). Another issue concerns data leakage (de Hond et al., 2022; Luo et al., 2016). The random-forest-based feature selection relied on all participants and their openness survey results, thus we cannot fully exclude that the test sample snooped into predictor space, though we consider this an unlikely explanation for the present results. Therefore, future feature selection should be based on the training and validation set.

As an alternative to web-search-based ICs, web tracking ICs could be considered. While cookies collect more information over a shorter time (e.g., Yan et al., 2022), it would be interesting to see whether the predictions become better with web tracking data.

6. Ethical considerations and data availability

As web tracking and web search data are already used for commercial applications, we feel that it is an ethical necessity to lead an open scientific discussion about the possibilities and limitations that come with such data. In contrast to these commercial objectives, we here aim to improve future psychodiagnostics and thus set the ground for computational methods improving psychotherapies.

We invested a considerable amount of time into the anonymization of the ICs. They do not contain client information, URLs or web site visiting time information and we keep only very coarse time stamps relative to the time of the assessment – thus identifiability of the participants should be low (Deußer et al., 2020). Moreover, our ICs can be considered as a subset of the information that is collected during web tracking. Therefore, identifiability of the participants should be lower than with standard web tracking. Nevertheless, we feel that de-anonymization hackatons on informed participants would be useful to test for the success of anonymization. Of course, we obtained an ethics committee approval for the present study and also thoroughly documented the data protection infrastructure for our own scientific use. As soon as data re-use by a greater research community is intended, however, de-anonymization could become more problematic. Participants already agreed that the data can be shared for scientific purposes, but at present we would hesitate to share the data, even as soon as the complete data collection is finished. Unsuccessfully testing for anonymization by hackatons would provide evidence that a legally penalizable non-disclosure agreement may be sufficient for scientific re-use. But at present, we hope that data centers will become available soon, in which computations over the data can be performed, while access to the raw data is strictly limited.

If any reader is interested in a shared task challenge for predicting psychological traits and/or for a hackaton examining de-anonymization, please contact the first author.

7. References

Ackerman, P. L. (1996). A Theory of Adult Intellectual Development: Process, Personality, Interests, and Knowledge. *Intelligence*, *22*, 227–257.

Anglim, J., Horwood, S., Smilie, L. D., Marrero, R. J. Wood, J. K. (2020). Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychological Bulletin, 146*(4), 279-323.

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159.

Biemann, C., & Riedl, M. (2013). Text: now in 2D! A framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, *1*(1), 55–95.

Burden, F. R., Ford, M. G., Whitley, D. C., & Winkler, D. A. (2000). Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Neural Networks. *Journal of Chemical Information and Computer Sciences*, *40*(6), 1423–1430.

Cattell, R. B. (1987). *Abilities: Their structure, growth, and action*. North-Holland.

Chao, W.-L. (2011). *Machine Learning Tutorial*. http://disp.ee.ntu.edu.tw/~pujols/Machine%20Learning%20Tutorial.pdf

de Hond, A. A. H., Leeuwenberg, A. M., Hooft, L., Kant, I. M. J., Nijman, S. W. J., van Os, H. J. A., Aardoom, J. J., Debray, T. P. A., Schuit, E., van Smeden, M., Reitsma, J. B., Steyerberg, E. W., Chavannes, N. H., & Moons, K. G. M. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *Digital Medicine*, *5*(1), 1–13.

Deußer, C., Passmann, S., & Strufe, T. (2020). Browsing unicity: On the limits of anonymizing web tracking data. *Proceedings - IEEE Symposium on Security and Privacy*, *2020-May*, 777–790. https://doi.org/10.1109/SP40000.2020.00018

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E. P., & Ungar, L. H. (2021). Closed and Open-Vocabulary Approaches to Text Analysis: A Review, Quantitative Comparison, and Recommendations. *Psychological Methods*, *26*(4), 398–427.

Ermakova, T., Fabian, B., Leipzig, H., Bender, B., & Klimek, K. (2018). Web Tracking – A Literature Review on the State of Research. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 4732–4741. http://hdl.handle.net/10125/50485

Goldberg, L. R. (1993). The Structure of Phenotypic Personality Traits. *American Psychologist*, *48*(1), 26–34.

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., & Jacobs, A. M. (2018). Simple Co-Occurrence Statistics Reproducibly Predict Association Ratings. *Cognitive Science*, *42*(7), 2287–2312.

Hofmann, M. J., & Jacobs, A. M. (2014). Interactive activation and competition models and semantic context: From behavioral to brain data. *Neuroscience and Biobehavioral Reviews*, *46*, 85–104.

Hofmann, M. J., Jansen, M., Wigbels, C., & Jacobs, A. M. (2023, March 28). *Individual internet search history predicts openness, interest, knowledge and intelligence*. Https://Www.Researchgate.Net/Publication/369559653_Individual_internet_search_history_predicts_openness_interest_knowledge_and_intelligence.

Hofmann, M. J., Müller, L., Rölke, A., Radach, R., & Biemann, C. (2020). Individual corpora predict fast memory retrieval during reading. *Conference: Proceedings of the 6th Workshop on Cognitive Aspects of the Lexicon (CogALex-VI)*, 1–10. https://www.researchgate.net/publication/344693965_Individual_corpora_predict_fast_memory_retrieval_during_reading

Hofmann, M. J., Remus, S., Biemann, C., Radach, R., & Kuchinke, L. (2022). Language Models Explain Word Reading Times Better Than Empirical Predictability. *Frontiers in Artificial Intelligence*, *4*, 1–20.

Holland, J. L. (1959). A Theory of Vocational Choice. *Journal of Counseling Psychology*, *6*(1), 35–45.

Jacobs, A. M. (2019). Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics. *Frontiers in Robotics and AI, 6*(53), 1-13.

Jacobs, A. M. (2023). *Neurocomputational Poetics: How the Brain Processes Verbal Art.* Anthem Press.

Jacobs, A. M., & Kinder, A. (2022). Computational Models of Readers' Apperceptive Mass. *Frontiers in Artificial Intelligence*, *5*(718690), 1–16.

Jankowsky, K., Krakau, L., Schroeders, U., Zwerenz, R., & Beutel, M. E. (2023). Predicting Treatment Response Using Machine Learning: A Registered Report (Stage 2). *Registered Reports*. https://www.researchgate.net/profile/Kristin-Jankowsky/publication/376645900_Predicting_treatment_response_using_machine_learning_A_registered_report/links/65820e0e0bb2c7472bf8884e/Predicting-treatment-response-using-machine-learning-A-registered-report.pdf

Johannßen, D., Biemann, C., & Scheffer, D. (2019). Reviving a psychometric measure: Classification and prediction of the Operant Motive Test. *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 121–125. https://aclanthology.org/W19-3014.pdf

Kandler, C., & Piepenburg, A. (2020). Leisure Interests and Engagement: Distinct Dispositions or only Expressions of Personality Traits? *Journal of Individual Differences*, *41*(2), 101–109.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805.

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., Derks, E. M., Fleischhacker, W. W., & Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry*, *3*(10), 935–946.

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., Ho, T. B., Venkatesh, S., & Berk, M. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research*, *18*(12), 1–10.

Lyon, K.A., Elliott, R., Ware, K., Juhasz G., Brown, L.J.E. (2021). Associations between Facets and Aspects of Big Five Personality and Affective Disorders: A Systematic Review and Best Evidence Synthesis. *Journal of Affective Disorders, 288*, 175-188.

McCrae, R. R., & Costa, P. T. (1987). Validation of the Five-Factor Model of Personality Across Instruments and Observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90.

Mehta, P., Bukov, M., Wang, C. H., Day, A. G. R., Richardson, C., Fisher, C. K., & Schwab, D. J. (2019). A high-bias, low-variance introduction to Machine Learning for physicists. *Physics Reports*, *810*, 1–124.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv Preprint ArXiv:1301.3781*, *1301.3781*, 1–12. http://arxiv.org/abs/1301.3781

Mussel, P., & Palaecke, M. (1996). BFAS-G. Big Five Aspect Scales - German. In Elektronisches Testarchiv (Ed.), *Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)*. https://doi.org/10.23668/psycharchives.2341

Nikstat, A., Höft, A., Lehnhardt, J., Hofmann, S., & Kandler, C. (2018). Entwicklung und Validierung einer Kurzversion des Fragebogeninventars für Freizeitinteressen (FIFI-K). *Diagnostica*, *64*(1), 14–25.

Ostendorf, F. (1990). *Sprache und Persönlichkeitsstruktur: zur Validität des Fünf-Faktoren-Modells der Persönlichkeit* (Psychologie; Bd.42). S. Roderer.

Ouyang, B., Song, Y., Li, Y., Wu, F., Yu, H., Wang, Y., Yin, Z., Luo, X., Sant, G., & Bauchy, M. (2021). Using machine learning to predict concrete's strength: Learning from small datasets. *Engineering Research Express*, *3*, 1–10.

Pennebaker, J. W., & King, L. A. (1999). Linguistic Styles: Language Use as an individual Difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312.

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of LREC 2010, Workshop New Challenges for NLP Frameworks*, 46–50. http://numdam.org

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The Power of Personality The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes. *Perspectives on Psychological Science, 2*(4), 313-345.

Rolfhus, E. L., & Ackerman, P. L. (1996). Self-Report Knowledge: At the Crossroads of Ability, Interest, and Personality. *Journal of Educational Psychology*, *88*(1), 174–188.

Satow, L. (2017). *Allgemeiner Intelligenz-Test (AIT)*. www.psychomeda.de/online-tests/

Schaumlöffel, L., Hübner, R., Thiel, S., & Stulle, K. P. (2018). Du bist, was du sprichst – Validierung der Sprachanalysetechnologie PRECIRE® anhand des HEXACO®-Persönlichkeitsmodells. In K. P. Stulle (Ed.), *Psychologische Diagnostik durch Sprachanalyse* (pp. 57–158). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-18771-2_3

Schipolowski, S., Wilhelm, O., Schroeders, U., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2013). BEFKI GC-K – A Short Scale for the Measurement of Crystallized Intelligence. *Methoden, Daten, Analysen* , *7*(2), 153–181.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, *8*(9), 1–16.

Viering, T., & Loog, M. (2021). The Shape of Learning Curves: a Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7799–7819. http://arxiv.org/abs/2103.10948

von Stumm, S., & Ackerman, P. L. (2013). Investment and intellect: A review and meta-analysis. *Psychological Bulletin*, *139*(4), 841–869.

Watrin, L., Schroeders, U., & Wilhelm, O. (2021). Structural Invariance of Declarative Knowledge Across the Adult Lifespan. *Psychology and Aging*, *37*(3), 283–297.

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2015). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *Quarterly Journal of Experimental Psychology*, *68*(8), 1599–1622.

Winter, D. G. (2010). Why achievement motivation predicts success in business but failure in politics: The importance of personal control. *Journal of Personality*, *78*(6), 1637–1668.

Yan, P., Schroeder, R., & Stier, S. (2022). Is There a Link between Climate Change Scepticism and Populism? An Analysis of Web Tracking and Survey Data from Europe and the US. *Information, Communication & Society*, *25*(10), 1400–1439.

Yarkoni, T. (2010). Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363–373.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(4), 1036–1040.

Ziegler, M., Danay, E., Heene, M., Asendorpf, J., & Bühner, M. (2012). Openness, fluid intelligence, and crystallized intelligence: Toward an integrative model. *Journal of Research in Personality*, *46*(2), 173–183.

Ziegler, M., Schroeter, T. A., Lüdtke, O., & Roemer, L. (2018). The enriching interplay between openness and interest: A theoretical elaboration of the OFCI model and a first empirical test. *Journal of Intelligence*, *6*(3), 1–22.

An Empirical Study on Vague Deictic Temporal Adverbials

Svenja Kenneweg[1], **Brendan Balcerak Jackson**[1], **Jörg Deigmöller**[2], **Julian Eggert**[2], **Philipp Cimiano**[1]
[1]Bielefeld University, Germany
{skenneweg, bbalcerakjackson, cimiano}@techfak.uni-bielefeld.de
[2]Honda Research Institute Europe, Germany
{Joerg.Deigmoeller, Julian.Eggert}@honda-ri.de

Abstract

Temporal adverbial phrases such as *recently* and *some time ago* have a special function in communication and temporal cognition. These adverbials are deictic, in that their meaning is tied to their time of utterance; and they are vague, in that the time periods to which they apply are under-specified in comparison to expressions such as *yesterday*, which precisely indicates the day before the day of utterance. Despite their vagueness, conversational participants have a mental image of when events described using these adverbials take place. Our study aims to quantify this mental model in terms of fuzzy or graded membership. To achieve this, we investigated the four English temporal adverbials *recently*, *just*, *some time ago* and *long time ago* as applied to types of events with different durations and frequencies, by conducting surveys to measure how speakers judge the different adverbials to apply in different time ranges. Our results suggest that it is possible to represent the meanings of deictic vague temporal adverbials geometrically in terms of graded membership within a temporal conceptual space.

Keywords: Temporal adverbials, Meaning representation, Deixis, Vagueness, Cognitive semantics

1. Introduction

Temporal adverbial phrases such as *recently*, *soon*, *some time ago* and *in a while* play a distinctive role in communication and cognition pertaining to time. For example,

(1) I just had my birthday.

If person A utters (1) to person B, B is likely to assume that A's birthday was in the past several days, and might respond by congratulating A. But suppose A utters the following instead:

(2) My birthday was some time ago.

B is unlikely to congratulate A, and B would be surprised to learn that A's birthday was in fact only two days prior.

Temporal adverbials like these are instances of *deixis*, or context-sensitivity: the range of times to which the adverbial applies depends on the time at which it is uttered, and different utterances apply to different ranges of times. Other deictic expressions are pronouns such as *I* and *you*, demonstratives such as *this* and *that*, and locative expressions such as *here* and *there*.

These adverbials are also instances of *vagueness*. For a given utterance of (1), for example, there is a range of times prior to the time of utterance which would not clearly be correctly described as times when the speaker had 'just' had her birthday, but which would also not clearly be described as time when she had *not* 'just' had her birthday. Other examples of vague expressions are gradable adjectives such as *tall* and *old*, adverbs such as *quickly* and *loudly*, and nouns such as *pile* and *heap*.

Not all vague temporal adverbials are deictic, e.g. *just before Christmas 2023*. And not all deictic temporal adverbials are vague, e.g. *exactly 37 minutes ago*. Our focus here is on temporal adverbials that are both deictic and vague. We take a cognitive approach, seeking to understand how the mind conceptualizes times and events in terms of vague deictic temporal adverbials.

We present an initial empirical study conducted with 100 adult participants concerning the past temporal adverbials *just*, *recently*, *some time ago*, and *long time ago*. For each adverbial, subjects were given a range of scenarios involving past events and were asked to evaluate the extent to which the expression applies in the scenarios. The aim of the study was to measure how language users represent events, both in terms of the time of utterance (deictic aspect) and in terms of fuzzy membership (vague aspect). In future work this study can lay the foundation in developing a computational model of vague temporal adverbials.

2. Related Work

2.1. Vagueness

Vagueness is standardly characterized by the existence of borderline cases: an expression is vague just in case there are cases in which ordinary speakers judge that the expression neither clearly applies nor clearly fails to apply. The

The Workshop on Cognitive Aspects of the Lexicon, pages 26–31
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

adjective ‘tall’, as applied to persons, counts as vague because there are borderline cases of tall people, i.e. people who neither clearly count as tall nor clearly count as not tall. Numerous logico-linguistic frameworks have been proposed for making sense of borderline cases and vagueness; useful surveys are provided in Keefe and Smith (1997); Keefe (2000); Kennedy (2019); Solt (2015).

The most directly relevant framework for the present study is the *fuzzy* approach to vagueness (Damerau (1977); Hájek (1998); Zadeh (1965, 1973)). While classical semantics treats membership in a class such as *tall people* as an all-or-nothing matter, fuzzy approaches distinguish different degrees of membership on the closed interval [0,1]. A borderline case of a tall person is someone whose degree of membership measures somewhere in the middle of this interval. Early fuzzy approaches tended to interpret graded membership extensionally as a relationship between the expression and entities in extra-linguistic reality (e.g. actual and possible people of different heights). Here we construe graded membership in cognitive terms, as representing the way language users mentally represent reality (see Hersh and Caramazza (1976); Douven et al. (2013)); we discuss this further in Section 2.2.

The vast majority of literature on vagueness either abstracts from specific categories of expressions, or else focuses on vagueness in the adjectival domain. Very little work focuses specifically on vagueness in temporal adverbials, and virtually none we could find investigates vague temporal adverbials experimentally. A notable exception is Van Jaarsveld and Schreuder (1985), an empirical study of temporal adverbials in Dutch. Their findings suggest that the range of times for which a speaker is disposed to apply a temporal adverbial to an event is influenced by the subject’s beliefs about the frequency and duration of events of that type. This study and its methodology provided the starting point for the present work.

2.2. Conceptual Space Semantics

Here we adopt a cognitive perspective on semantics that takes linguistic expressions to correspond to concepts or ways of mentally representing reality. Particularly relevant for the present study is the *conceptual spaces* framework of Gärdenfors (2014), in which concepts are represented geometrically, as regions in spaces that are defined in terms of one or more representational dimensions. For example, humans represent color using a three-dimensional conceptual space defined in terms of the dimensions *hue*, *saturation*, and *brightness*. Color concepts, and the meanings of color terms, correspond to regions of this conceptual space that have certain formally specified properties (Gärdenfors (2014)); the location of a given object within a region indicates the way the subject represents its color. Other concepts correspond to regions in conceptual spaces constructed from dimensions corresponding to other qualities, such as spatial or temporal extent, auditory experience or taste, and different kinds of motion and action.

As Douven et al. (2013) and Decock and Douven (2014) argue, conceptual spaces provide a natural way of interpreting the graded membership relation employed by the fuzzy approach to vagueness. The metric of a conceptual space makes it possible to identify *prototypes* for a concept, subregions that have distinguished positions within the region corresponding to the concept. The degree to which an object belongs to the concept can then be calculated using its distance from the prototypes for the concept together with its distance from the prototypes of adjacent concepts in the space. A borderline case of *red*, for example, might have a degree of membership of 0.4 because it is just about as far away from the prototypes for red as it is from the prototypes for orange. The mathematical details can become somewhat complex, especially for high-dimensional conceptual spaces (see Decock and Douven (2014) for discussion) but they are not necessary for present purposes.

3. Experimental Setup

We devised a set of online surveys to measure language users’ representations of events in terms of a graded or fuzzy relation to the time of utterance. Our study centered on the four temporal adverbials *recently*, *just*, *long time ago* and *some time ago* as applied to hypothetical events involving an imaginary person named Tom. In light of the results in Van Jaarsveld and Schreuder (1985), which suggest that the applicability of a temporal adverbial is influenced by the duration and frequency of the associated event, we designed five separate surveys for five types of events with different durations and frequencies: Birthday, Brushing Teeth, Marriage, Vacation and Year Abroad. Brushing Teeth has the shortest duration, which is typically only a few minutes. In contrast, Year Abroad has a significantly longer duration of one year. The durations of the other events fall somewhere in between these extremes. The frequency of events follows a comparable reversed hierarchy: Brushing Teeth happens daily and has therefore the highest frequency, while Year Abroad happens only a few times, if at all, in one’s life. The frequencies of the other events lie in between.

Each survey queried subjects about a series of English sentences containing a temporal adverbial

	Just *Recently*	*Some time ago* *Long time ago*
Brushing Teeth	5 min. ... 1 day	1 hour ... 3 days
Birthday	1 day ... 3 months	1 week ... 11 months and 3 weeks
Vacation	1 day ... 4 months	4 days ... 1 year
Marriage Year Abroad	4 days ... 1 year	1 month ... 8 years

Table 1: The nearest and furthest $< timeSpan >$ point surveyed for each event type and each temporal adverbial.

applied to a past event (see section 3.1). Participants were asked to rate the applicability of the temporal adverbial in each case on a 5-point Likert scale; ”Doesn't apply”; ”Barely applies”; ”Partially applies”; ”Mostly applies”; ”Completely applies”. Each survey was administered to 100 participants via the Prolific platform, and each participant received a small compensation for completing the survey.[1] All participants were adults with English as their native and primary language.

3.1. Survey sentences

Each of the surveys had one survey page for each of the four temporal adverbials, with 7 test sentences on each survey page. Test sentences were constructed using an event description and a $< timeSpan >$ indicating the time passed since the occurrence of the event, followed by a statement applying the temporal adverbial to the event. For example, for the event type Birthday, the $< timeSpan >$ 1 day and temporal adverbial *recently*, the test sentence is: ”Tom's Birthday was 1 day ago. Statement: Tom had his Birthday recently.”

The range of values for $< timeSpan >$ was chosen separately for each type of event based on our own experience. For each event type we maintained one range of $< timeSpan >$ for *just* and *recently*, which concern times closer to the present, and a different range for *some time ago* and *long time ago*, which concern more distant time points. The range for $< timeSpan >$ for each temporal adverbial and event is shown in Table 1.

[1]https://www.prolific.com/

4. Results

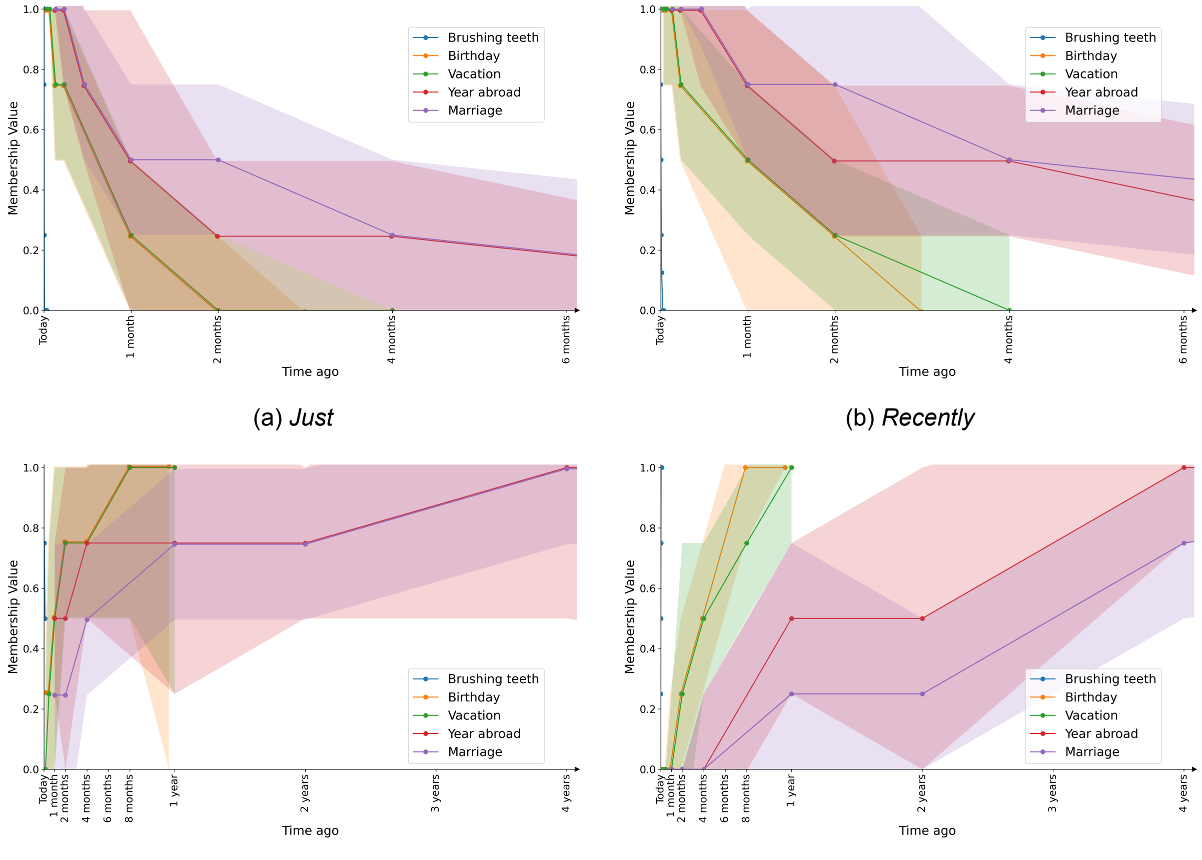
Figure 1 shows the median and interquartile ranges (highlighted in brighter colors) of all survey data for each of the temporal adverbials studied, with all five event types plotted together for comparison. The Likert scale points from ”Doesn't apply” to ”Completely applies” were numerically encoded from 1 to 5. Subsequently we applied a linear transformation to map the values onto a range of 0 to 1, representing degree of membership.

We first compare the plots for the different temporal adverbials, starting with a comparison of *just* and *recently* (Figures 1(a)-(b)). Membership values for both adverbials are highest when the event has just occurred and decrease as the distance into the past increases. However, the decrease is more pronounced for *just* than *recently* across all event types. For example, although the membership value for a marriage that took place four days ago is 1 for both adverbials, speakers judge *recently* to apply to a marriage that took place 4 months ago more strongly than *just*. Turning to *some time ago* and *long time ago* (Figures 1(c)-(d)), we see that the initial values for *some time ago* are higher than those for *long time ago*. For Brushing teeth (which is hidden by the Time Axis) and *some time ago* the values rise to a peak before falling off again as we move further into the past. Our data does not show such a peak for the other events but it is possible that it lies outside the range of times we surveyed. An indication of this is that the interquartile range of Birthday and Vacation is higher at the end for *some time ago*. This suggests that *some time ago* corresponds to a segment of the past with vaguely defined start and end points, while *long time ago* corresponds to any time further in the past than a vaguely defined starting point.

When we compare the different event types across all temporal adverbials, we can see that the arrangement of the event types is always the same: first Brushing Teeth, then Birthday, Vacation, and Year Abroad, and then finally Marriage. The rate of increase or decrease in membership values is always slowest for Marriage and fastest for Brushing Teeth.

5. Discussion

Our results have a natural interpretation within the conceptual spaces framework. Sinha and Gärdenfors (2014) hypothesize a conceptual space defined in terms of *D-time* (deictic time), which orders points in terms of their distance (forwards or backwards) from the present. The meanings of deictic adverbials can be represented as regions in this space, and our results can be understood as helping to map the contours of those regions. This also provides a natural framework for repre-



(a) *Just*

(b) *Recently*

(c) *Some time ago*

(d) *Long time ago*

Figure 1: Median survey data and their interquartile ranges (shown in brighter colors) for the 4 evaluated temporal adverbials and 5 events.

senting vagueness in temporal adverbials in terms of graded membership, as sketched in Section 2.2. Gärdenfors (2014) conjectures that the meanings of (nearly) all simple natural language expressions correspond to convex regions of conceptual space. Within a graded membership framework, convexity requires that all the points on the path between any two points that belong to the region to at least membership value α also belong to the region to at least value α. All the regions of D-time measured are convex in this sense.

Our findings are also consistent with the earlier study of Van Jaarsveld and Schreuder (1985) that the interpretation of a temporal adverbial is influenced by the duration and frequency of the event. For each adverbial, all five event types have the same curve progression. (The time scale in Figure 1 makes this difficult to see for Brushing Teeth.) This suggests that the primary effect of combining an adverbial with different event types is to stretch or compress the corresponding region of conceptual space, rather than to produce significant changes in its shape.

Moreover, the methodology used here establishes natural hierarchies among times according to whether a given temporal adverbial, combined with a given event type, applies more clearly or definitely at one time or another. In principle, this can put us in a position to determine the most probable time of occurrence of an event given its linguistic representation in terms of temporal adverbials; meaning representations that support this kind of inference would be especially valuable for computational implementations.

6. Conclusion & Future Work

We presented a study that measured language users' representations of events in terms of a graded relation to the time of utterance, in order to simultaneously capture the deictic and vague character of the temporal adverbials selected. Our meaningful results provide confirmation that the survey method we employed is suitable for this purpose. Our results are consistent with the work of Van Jaarsveld and Schreuder (1985), and they have a natural interpretation within the conceptual spaces semantic framework.

Our eventual aim is to develop a computational model for applications such as human-robot interaction. Given an event type and a temporal adverbial, such a system should be able to recover a time span in which the event is most clearly represented as occurring; it should also be able to choose appropriate temporal adverbials for communicating temporal information. Reaching this point requires much more work. A first step is determining exactly how event duration and frequency affect temporal adverbial distribution over deictic time, an additional step being to explore other influencing factors. Other possible influencing factors can be identified by comparing events with the same durations and frequencies, to check for differences in mental representations of the corresponding time spans that cannot be entirely accounted for in terms of the influence of duration and frequency.

Given the recent progress in large language models, another important next step is to check whether they interpret vague deictic temporal adverbials in a way that is consistent with human understanding. This will help determine what role large language models can play in computational modeling of temporal communication and cognition.

This paper focused on deictic temporal adverbials and events. Another area of focus for future work could be on relative adverbials that relate one event to another, such as *Tom was on vacation just before his birthday*.

7. Optional Supplementary Materials

7.1. Limitations

The survey was limited by the number of questions that participants would be likely to answer seriously. According to Bowling et al. (2021), the proportion of unfocused random answers in online studies is less than 1% for studies with no more than 33 questions. Per survey, there are 28 questions about the vague temporal adverbials, two demographic questions and one question about the survey difficulty. Due to budget constraints, there were limits to the number of survey participants and the number of surveys that could be carried out.

7.2. Ethics

An ethics proposal for the surveys was submitted to the ethics committee of the Bielefeld University and was found to be ethically unobjectionable. This complies with the ethical guidelines of the German Psychological Association and the Professional Association of German Psychologists.

8. Bibliographical References

Nathan A. Bowling, Anthony M. Gibson, Joseph W. Houpt, and Cheyna K. Brower. 2021. Will the questions ever end? person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, 24(4):718–738.

Fred J Damerau. 1977. On “fuzzy” adjectives.

Lieven Decock and Igor Douven. 2014. What is graded membership? *Noûs*, 48(4):653–682.

Igor Douven, Lieven Decock, Richard Dietz, and Paul Égré. 2013. Vagueness: A conceptual spaces approach. *Journal of Philosophical Logic*, 42:137–160.

Peter Gärdenfors. 2014. *The geometry of meaning: Semantics based on conceptual spaces*. MIT press.

Petr Hájek. 1998. *Metamathematics of fuzzy logic*. Springer.

James A Hampton. 2007. Typicality, graded membership, and vagueness. *Cognitive Science*, 31(3):355–384.

Harry M Hersh and Alfonso Caramazza. 1976. A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General*, 105(3):254.

Rosanna Keefe. 2000. *Theories of vagueness*. Cambridge University Press.

Rosanna Keefe and Peter Smith. 1997. Introduction: theories of vagueness. In Rosanna Keefe and Peter Smith, editors, *Vagueness: a reader*. MIT press.

Christopher Kennedy. 2019. Ambiguity and vagueness: An overview. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: lexical structures and adjectives*. Walter de Gruyter.

Charles I. Mosier. 1941. A psychometric study of meaning. *The Journal of Social Psychology*, 13(1):123–140.

Terence Parsons. 1990. Events in the semantics of english: A study in subatomic semantics.

Chris Sinha and Peter Gärdenfors. 2014. Time, space, and events in language and cognition: a comparative view. *Annals of the New York Academy of Sciences*, 1326(1):72–81.

Stephanie Solt. 2015. Vagueness and imprecision: Empirical foundations. *Annu. Rev. Linguist.*, 1(1):107–127.

H.J. Van Jaarsveld and R. Schreuder. 1985. Implicit quantification of temporal adverbials. *Journal of Semantics*, 4:327–339.

Frank Vlach. 1993. Temporal adverbials, tenses, and the perfect. *Linguistics and Philosophy*, 16(3):231–283.

Lofti Zadeh. 1965. Fuzzy sets. *Information and Control*, 8:338–353.

Lotfi Zadeh. 1973. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on systems, Man, and Cybernetics*, (1):28–44.

Symbolic Learning of Rules for Semantic Relation Types Identification in French Genitive Postnominal Prepositional Phrases

Hani Guenoune, Mathieu Lafourcade
University of Montpellier, LIRMM
161 Rue Ada, 34095 Montpellier, FRANCE
{hani.guenoune, mathieu.lafourcade}@lirmm.fr

Abstract

We are interested in the semantic relations conveyed by polylexical entities in the postnominal prepositional noun phrases form "*A de B*" (A of B). After identifying a relevant set of semantic relations types, we proceed, using generative AI, to build a collection of phrases, for each semantic relation type identified. We propose an algorithm for creating rules that allow the selection of the relation between A and B in noun phrases of each type. These rules correspond to selecting from a knowledge base the appropriate neighborhood of a given term. For the phrase "*désert d'Algérie*" carrying the *location* relation, the term "*désert*" is identified as a geographical location, and "*Algérie*" as a country. These constraints are used to automatically learn a set of rules for selecting the *location* relation for this type of example. Rules are not exclusive as there may be instances that fall under multiple relations. In the phrase "*portrait de sa mère - the portrait of his/her mother*", all of *depiction*, *possession*, and *producer* types are a possible match.

Keywords: Genitive, Postnominal prepositional noun phrases, Semantic relations

1. Introduction

Beyond the necessity of identifying polylexical entities for automated language analysis, it is important for various applications to also understand the nature of relation binding the different components of polylexical terms. Our focus is on the genitive case of the nominal complement "de N" in French. In other words, compound noun phrases (NP) formed through the use of the preposition "*de*" introducing a syntactic complement B to a head A in "A de B", where A and B are nouns (and variations "*A d'B*", "*A du B*", etc.). We aim to automatically identify the semantic relation between terms A and B. Generally, such an approach contributes to a richer interpretation of discourse in textual content and leads to better semantic representations. Among the tasks benefiting from this study, we can mention the Question Answering task (Kapanipathi et al., 2020; Ben Abacha, 2012), which requires a rich semantic representation of text and relations between mentioned entities. Another task is the resolution of anaphors triggered by possessive determiners which involves the transformation of genitive forms into anaphoric phrases ("*le vélo de Julie* $\rightarrow$ *son vélo*"), and in which the resolution is based on constraints of the relation type between the anaphor and its antecedent (Guenoune, 2022).

In our project's specific context, these efforts also contribute to consolidating a common-sense knowledge base, firstly through the identification of semantic relation types relevant to our various inference mechanisms and Natural Language Processing (NLP) applications that leverage the knowledge base. Another way of improving the quality of the knowledge base is to develop a classification system that serves as a control tool. The analysis of the correctness of the results of such a tool would bring insights into the overall quality of the knowledge used. In order for this to be possible, an emphasis has to be put on the explainability of the methods to use, at is essential in identifying potential gaps and highlighting appropriate ways of consolidating the knowledge base.

In "A de B," the nominal head (A) plays a crucial role in maintaining the underlying sense with its complement (B). The type of nouns linked by the preposition in a genitive construction conditions the semantic relations that bind them. Regarding nominal possessives such as "*John's friend*" (which translate to a "A de B" NP in French (*l'ami de John*), the distinction mentioned in (Barker et al., 2019) differentiates the use of *sortal* nouns from (two-place) *relational* nouns. The contrast in their definition is analogous to that between *unary* and *binary* predicates in first-order logic. The *relationality* of a noun (De Bruin and Scha, 1988) concerns whether a referent must be identified in relation to another entity. In constructing nominal phrases, relational nouns only make sense when related to exactly two arguments, as seen in the example of *family/kinship* names *[father, mother, sister, brother...]* considered as archetypal relational nouns (see Löbner classification system (Löbner, 2011)).

The semantic relation in the phrase "*la mère de Lucie - the mother of Lucie*" relates solely to the sens of the nominal head (the *relational* noun "*mère - mother*"). Thus, a genitive noun phrase with a *relational* nominal head allows for a *lexical interpretation* of the semantic relation (De Bruin and Scha, 1988), contrasting with pragmatic readings that re-

The Workshop on Cognitive Aspects of the Lexicon, pages 32–41
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

sult from the use of certain *sortal nouns*, requiring extra-lexical information to identify the nature of the semantic link in postnominal NPs. Mentioning "*le nuage de Lucie - Lucie's cloud*" requires the introduction of pragmatic elements to fully grasp the type of relation (*the cloud she was looking at, drawing, dreaming of, etc.*).

The meanings conveyed by this type of NPs are therefore diverse, even though automatic interpretation efforts often reduce the types of semantic relations to one of member-collection/possession types (one explanatory element for this simplification could be the importance of theses relations and their predominant role in the standard semantic typologies considered in NLP works).

Beyond the typological framework of nominal heads, specific nouns (whether relational or sortal) introduce a multitude of possible relations between the terms of the NP. This work aims to study the nature of semantic links in this configuration and proposes a semantic typology for these links. Furthermore, we introduce a symbolic learning algorithm which serves as a basis for an explainable system of classification of semantic types in genitive NPs. We note that since figurative meanings are revealed to be challenging to determine computationally, this work does not address the identification of the overall meaning of the form "*A de B*" when it has acquired an idiomatic/figurative sense (*e.g.* "*homme de paille / écran de fumée - straw man / smoke screen*") .

This paper covers several aspects:

1. Proposal of a typology of semantic relations in postnominal genitive phrases, followed by the creation of an associative corpus between examples of genitive constructions and corresponding relation types. Data are collected using a generative AI, cautiously validated by hand. A portion of the corpus serves as training data while the rest serves as a test set.

2. Introduction of *GRASP*-it, a learning algorithm calculating decision rules for probable relation types.

3. Evaluation of the quality of produced constraints by implementing a second algorithm for classifying semantic types in "A de B" forms.

We begin with an overview of the resources used in this project, namely the knowledge base that we seek to improve through this work, then the data used in developing *GRASP*-it. We also provide examples of each relation type resulting from the use of "de N" constructions. We then describe the learning mechanism implemented to synthesize semantic relations between terms of each type. Finally, we conduct an evaluation of the quality of the produced rules applied to a portion of the corpus.

2. Data

This project required the use of external resources to successfully carry out the study in general, the learning process, as well as the evaluation of the learned rules and the classification system.

2.1. Knowledge Base Used

The world knowledge that supports our study and the algorithms developed is built from the latest issue of the *JeuxDeMots* (*JDM*) project data collection (dated February 11, 2024) (Lafourcade, 2024).

JeuxDeMots (*JDM*) (Lafourcade, 2007) is a lexical-semantic network represented by a directed graph. Graph nodes represent terms, while arcs signify typed, weighted, and potentially annotated relations between terms. The graph tackles lexical polysemy by specifying hierarchical sense "*refinements*", where a specific sense is affiliated with the general sense of the term. *JDM* is based on practical tools, principles, and concepts (*e.g.* the notion of refinement, the diversity of semantic types, inverse relations, such as *r_isa* and *r_hypo*, and a series of collaborative tools). The *JDM* network is designed to be used as a knowledge support for AI solutions (semantic text analysis, reasoning, decision-making, automatic summarization, etc.). A weighting and symbolic valuation system (meta-information annotation, *e.g. rare, relevant, non-relevant*, etc.) has been implemented to facilitate graph traversal and its exploitation (Lafourcade and Le Brun, 2023). As of February 1, 2024, *JDM* contained approximately $540\ million$ relations between over $9\ million$ terms and $24\ million$ nodes.

One central challenge for this project is to enrich our knowledge base with semantic information, particularly information regarding relations in genitive prepositional noun phrases. This helps text analysis and knowledge extraction. Indeed, When encountering a genitive form in a text it is desirable to guess the relation(s) between A and B.

2.2. Corpus of Genitive Constructions

We present a small-sized corpus for the learning and evaluation of semantic type determination rules. This corpus is to be seen as a starting point for the creation of larger scale collections. Despite the significance of small corpora (Landragin, 2018), below several thousand examples, it reveals challenging to apply resource-intensive procedures such as neural learning algorithms. However, we aim to integrate this effort into a longer-term project where data augmentation procedures can be implemented, such as automatic semantic enrichment mechanisms or manual annotation completion.

In the following, we detail semantic types identified, then discuss the data acquisition and validation protocol. In order to avoid introducing any bias, we chose to collect data from a source independent of the *JDM* knowledge base.

2.2.1. Semantic Typology

In Table 1, we list the considered semantic types along with an explanation and examples, and the corresponding relation type in *JDM* (with the appropriate orientation, where relations with names in the form '*r_x-1*' denote the conversive relation of '*r_x*').

Relation Type	***JDM* Relation**
Consequence (Co): *Term A is a consequence of (caused by) term B.*	*r_has_causatif*
dégâts de la tempête - retards de la circulation *(EN) storm damage - traffic delays*	
Possession (Po): *A is possessed by B*	*r_own-1*
fusil du soldat - vélo du cycliste *(EN) the soldier's rifle - the cyclist's bike*	
Material/composition (M): *Term A is composed of or is made of material B.*	*r_objet>matière*
cuillère de bois - trône de fer *(EN) wooden spoon - iron throne*	
Origin (O): *Term A originates from the location B.*	*r_lieu>origine*
vin de France - café du Brésil *(EN) wine from France - Coffee from Brazil*	
Topic *(T)*: *Term A has term B as its theme (or subject).*	*r_topic*
restaurant de sushis - film d'horreur *(EN) sushi restaurant - horror movie*	
Quantification (Q): *A serves as a measure for B.*	*r_quantificateur*
brin d'herbe - minute d'attente *(EN) blade of grass - waiting minute*	
Depiction (D): *Term A is a depiction of term B.*	*r_depict*
peinture d'un paysage - photo d'une famille *(EN) painting of a landscape - photo of a family*	
Characterisation (Ca): *Term A is a property or the noun of an adjective that can qualify term B.*	*r_has_property-1*
sournoiserie du politicien - sagesse du viellard *(EN) the politician's cunning - the elder's wisdom*	
Holonymy(H): *Term A is part of term B.*	*r_holo*
coque du bateau - écaille du poisson *(EN) boat hull - fish scale*	
Location (L): *Term A can have term B as its location.*	*r_lieu*
tour de Pise - sahara d'Algérie *(EN) tower of Pisa - sahara of Algeria*	
Agent (A): *Term A is an action in which the actor is term B.*	*r_processus _agent*
travail de l'ouvrier - cours du professeur *(EN) worker's job - the teacher's class*	
Patient (P): *Term A is an action that term B undergoes.*	*r_processus _patient*
travail du bois - ouverture de la porte *(EN) woodworking - opening of the door.*	
Instrument (I): *Term A is an instrument of B or an action that B undergoes.*	*r_processus >instr-1*
clé d'ouverture - clé de la porte *(EN) opening key - key of the door*	
Kinship/Social tie (LS): *A plays a role of 'A' in relation to B.*	*r_social_tie*
avocat d'une femme battue - chef du groupe *(EN) lawyer for a battered woman - group leader*	
Producer (AC): *A is produced by B.*	*r_product_of*
portrait de Van Gogh - gâteau du pâtissier *(EN) Van Gogh's portrait - pastry chef's cake*	

Table 1: List of semantic types considered and correspondences with semantic relations types in the lexical-semantic network *JDM*.

It should be noted that this particular list is the one we have established as basis for our study. Choices regarding granularity and the number of types were made to align this typology with the requirements of the resources and tools we use, as well as the needs of the applications that will take advantage of this typology. The number and type of semantic relation is thus very knowledge-base directed and strictly semantically expressed rather than pragmatically expressed as was the case with previous examples for English documented in the literature as state of the art (Nastase and Szpakowicz, 2003). It is by no means an exhaustive list of all possible types of relations between terms A and B in a "A de B" form.

Some types may be added to this list. It is also possible to specify/generalize certain types so that they more or less precisely correspond to theoretical frameworks that are different from ours. This is especially relevant when another knowledge base (than *JDM*) is used, potentially one defining a different set of relation types. Among these types, we can mention examples involving absolute temporal semantic relations (carried by class names), such as *repas de midi - brise du matin - bus de nuit* (*midday meal - morning breeze - night bus*) or spatial and temporal relative links (carried by relational nouns) such as *milieu/droite/gauche de la pièce - bas de page* (*middle/right/left of the room - bottom of the page*). Another case is that of nominations: "*Théorème de Pythagore - Râle de Wallace - Kappa de Fleiss"* (*Pythagoras' theorem - Wallace's rail - Fleiss' kappa*) which could be the subject of a separate category. For the mentioned cases, we choose to include them in types of similar semantics; for example, the first two cases are included in the "topic" type, while the case of nominations is classified among instances of "author/creator" (even though it may not necessarily involve a creation *per se*).

2.2.2. Data Collection and Validation

For each type of semantic relation mentioned above, we employed a generative AI (LLM conversational agent[1]) to generate a set of examples. We limited each type to 80 examples, with 50 dedicated to *training* and 30 for *test* purposes.

The strategy for constructing queries to the conversational agent varied depending on the types of relations. Obtaining satisfactory examples proved more or less challenging, depending on the cases. For types where the generated examples were less exploitable, we chose to guide the model through examples. We provided about ten examples composed by us, then explained commonalities at the level of underlying semantic relations over several conversational turns with the chatbot.

Although this iterative approach yielded examples of the desired type, it had the drawback of excessively "influencing" the responses generated by the chatbot, resulting in a set of polylexical entities with low diversity (strong alignment with the examples presented to the agent). Consequently, for the sake of diversification, *after an initial generation of examples*, we emphasized the need to diversify the generation in subsequent queries. This strategy was repeated until we considered the set convincing. However, the generated instances still contained approximately 10% misclassified or duplicated examples and remained imperfect in terms of diversity. Therefore, we conducted a manual validation of all examples produced by the chatbot. Specifically, the validation involved replacing duplicated cases and very similar entities, as well as reclassifying misclassified examples.

2.2.3. Data Formatting and Post-Processing

The produced collection includes examples of variable morphology. Concerning the presence or absence of a determiner for the nominal complement (term B), one might assume that a morphological normalization step would be beneficial. However, this criterion constitutes a morpho-syntactic marker that can be crucial for classification, which is why we choose not to perform morphological or lexical transformations. Nevertheless, it should be noted that polysemy of corpus terms may be observed. Therefore, exploiting the corpus will require data preparation, specifically a phase of semantic disambiguation to select the appropriate senses of the terms.

3. Presentation of *GRASP*-it

The *GRASP*-it (*Genitive Relations And Semantic Pattern Identification Tool*) algorithm aims to produce a set of constraint pairs for each type of relation based on input data. These constraints are based on the semantic types of the nominal head and the complement. They can be considered a synthesis of semantic attributes regarding the content of a knowledge base. The purpose of this set of constraints is to guide a classification process of semantic relations in genitive NPs. Another objective is to produce "*interpretable*" constraints that can easily be read and explained. In general, the first step of *GRASP*-it involves storing, for each example of a certain type, semantic information that could allow to classify the example in the relevant type:

- *Hypernyms* of terms A and B: The goal is to capture, as precisely as possible, the semantic *"types"* of both terms. An hypernym is a *term* (lexical entity) in *JDM* attainable through the relation *r_isa*.
- *Target for Relation Types* (*TRT*): A selection of relation types leading to the term. For example, a term frequently targeted by the *location*

[1]ChatGPT. Model version gpt-4-0613. 2023-06-13

relation is considered, by this approach, as a location. This enables the reinforcement of the relevance of this semantic class for a specific term. The selection of relations leading to the terms can be viewed as a means to supplement the list of hypernyms for a given term.

- *The Standard Semantic Type* (*SST*): through the relation *_INFO-SEM*, the standard type associates a lexicalized term with a standard ontological (conceptual) type.

The result of this step is a set of weighted pairs, referred to here as *signatures* of terms A and B. The number of pairs at this stage corresponds to the number of examples for each type, which, in the case of our corpus (training portion), amounts to 50 NP units of the form "A de B."

A signature is defined as an unordered set of symbols. Each symbol takes a value of a specific entry of JDM. For example, the signature s associated to the term "*véhicule*" would be as follows.

```
s(véhicule) = {
véhicule, transport urbain, partie de l'espace, Transport urbain ,
mode de transport, instrument, lieu,
transport, moyen, machine, moyen de
transport,

r_isa, r_hypo, r_has_part, r_holo,
r_agent, r_patient, r_lieu, r_instr,
r_carac-1, r_lieu-1, r_action_lieu,
r_mater>object, r_processus>agent,
r_own, r_is_instance_of,

_INFO-SEM-SUBST, _INFO-SEM-THING-ARTEFACT, _INFO-SEM-PLACE, _INFO-SEM-THING-CONCRETE, _INFO-SEM-PLACE-HUMAN
}
```

For clarity's sake, we divided the symbols into three blocks: *Hypernyms*, *TRT*s and *SST*. It's worth noting that the signature of a term contains the term itself, this aims to capture instances that are hyponyms of the signed term.

In addition to the need for its explainability, this representation of terms is designed to be controllable in terms of its content and size. This allows the *GRASP*-it method to be adaptable to the variable requirements of the application for which it is used.

The second step aims to aggregate *rules* of each type to process the entire set by generalization. As shown in (1), we define a *rule* R as a pair of constraints s_L and s_R (which are signatures, respectively *left* and *right* corresponding to terms A and B) and a semantic relation type rt.

$$R :< s_L, s_R, rt > \quad (1)$$

The aggregation is a *fusion* operation of two rules and is defined in (2).

$$Fusion(R1, R2) =< s1_L \cup s2_L, s1_R \cup s2_R, rt >^{2} \quad (2)$$

A fusion of two rules means that the constraints they respectively associate are sufficiently *similar* to be represented by a single pair of constraints. Formally, as a signature s can been seen as a vector, we adopted the *cosine similarity* (dot product divided by the product of norms), denoted as sim. Two signatures are considered sufficiently similar when their sim value is above a threshold of 0.5 (which has been set empirically). The merged signature is the vector sum of the two signatures (which corresponds to the union set).

A pair produced by one or more successive fusions is considered more general and *reliable* than a pair that has not undergone fusion. Reliability is therefore a measure of coverage of examples of the type and is calculated by assigning a weight to the pair of constraints corresponding to the number of fusions performed to arrive at the final form of the pair. At the output of this step, a set of more or less aggregated pairs of constraints, with a cardinality at most twice the number of examples of the considered type, is assigned to each type of relation (listed in Table 1).

The idea behind merging rules is that the result of successive fusions is a rule that *represents* appropriately a large set of examples of a certain type. A merged rule can thus be considered as a generalised *model* for a given relation type. One relation type can be associated to several models. A good *model* will appropriately associate the relation type between two terms A and B in a genitive NP.

4. Evaluation of *GRASP*-it

In this section, we present the conditions under which our evaluation was conducted and conclude with the performance scores of our system.

4.1. Data Preparation

A minimal phase of data preparation has been undertook before applying the classification algorithm (detailed in section 4.2). We identified two main tasks that have to be done prior to classification in order to take the entirety of the corpus into account.

Compound Words Identification: instances of NPs containing multiple prepositions "*de*" as in "*lunettes de soleil de marque - détecteur de fumée de protection*" raise the problem of choosing the right separating preposition. This has a direct influence on the identification of the terms A and B,

[2] Only two rules with the same rt can be merged.

hence the appropriate relation type to be identified. We proceeded to identify these instances by verifying their existence in the knowledge base. In "*lunettes de soleil de marque*", the candidates for terms A and B would be "*(**A:** lunettes, **B:** soleil de marque)*" and "*(**A:** lunettes de soleil, **B:** marque)*". In the former, the nonexistence of term B in *JDM* allows us to assign the latter candidate's values to A and B. In case both candidates result in known A and B terms, the intended separating preposition is manually annotated.

Generic Named Entities: Instances containing named entities such as person's first or last names are only well represented in our knowledge base when these entities are renowned or a common knowledge in popular culture (e.g. "*Coca-Cola*" - "*Lucie*"). Therefore, it is important to take into account nonexistent instances by carrying out transformations that replace the name by another (of the same type) that we know is well represented in the knowledge base.

4.2. Application Algorithm

To evaluate the pairs of semantic constraints (rules) produced by the learning algorithm, we implement a validation process that seeks to verify the satisfaction of these constraints. The goal is to identify the semantic types derived from the portion of the corpus not involved in the constraints/rules calculation. This involves 450 examples distributed evenly across the 15 possible types of relations (30 each).

4.2.1. Decision Criteria

Conceptually, the constraints validation approach is based on searching a similarity of semantic types between the terms of the input phrase and the set of terms from which the *GRASP*-it system was trained. The idea is to *induce* an *identity* of types if a test's terms are sufficiently *similar* to the semantics synthesizing the inputs of the learning process.

In practice, the search is conducted by calculating a similarity between terms A and B of the input and the respective signature in all the rules $< s_L, s_R, rt >_i$ learned by *GRASP*-it. The two obtained similarities (for term A with s_L and B with s_R) are aggregated by an arithmetic mean. Therefore, the similarity between a form "A de B" and a rule (pair of constraints) $< s_L, s_R, rt >$ is given in (3).

$$^1/_2(sim(s(A), s_L) + sim(s(B), s_R)) \tag{3}$$

One should note that a *signature* for a term and a *constraint* of a rule share an identical structure.

A positive response is returned for the type best ranked in terms of average similarity values with each pair. Note that for the ranking to be done, the verification procedure is carried out for all types and calculated pairs of constraints (once the training, and all possible fusions have been performed for all possible types).

It remains possible to only use the rules that are either the result of a fusion, or that have not been merged. That is to say, we could exclude rules that have been merged into a new rule. We expect that this *trimming* of the set of rules would lead to a shorter execution time without highly degrading the results (this particular point is subject of a study, see experiment 3).

4.2.2. Extra-Semantic Features

The detection of certain types depends more or less on extra-semantic markers, such as the use or not of a determiner, the use of named entities, or the definiteness of nominal complements. An example illustrating this is the phrase "*photo de famille*" as opposed to "*photo d'une famille*", in which the presence or absence of the determiner, conditions the interpretation of the semantic link (*topic* or *depiction*, *respectively*).

Such heuristics are not part of the core component of our solution, as we tackle the problem of highlighting the implication of *semantic* rules especially. Nevertheless, since the representation of terms is as simple as a set of symbols, the solution proposed seems also suitable for taking into account extra-semantic information. This amounts to the inclusion of the relevant traits into the signatures. Therefore, we proceeded to integrate the *definiteness trait* of complements (B) to our representations. This constitutes a separate study in the subsequent section (see experiment 2).

4.3. Evaluation Protocol

In this evaluation, we will conduct three distinct experiments, aiming to assess the following aspects.

Experiment 1: We seek to assess individually the quality gained from every semantic trait included in the signatures. Our baseline in this experiment is the use of *Hypernyms* only. The idea is to proceed contrastively and analyse the cases that we could successfully classify by adding separately the TRT, and SST traits.

Experiment 2: The inclusion of morphological markers could reveal beneficial to the classification process. Without dedicating *ad hoc* elaborate heuristics for these traits, we experiment the effects of constructing signatures with non-semantic traits, namely the *definiteness* trait. We use a straightforward symbolic approach discussed in 4.4.2.

Experiment 3: Here, we are interested in estimating the overall value for additional computational effort (which translates into a cost in execution time for example). The developed system is also to be used as a component of specific *NLP* tasks (the most important among them being relation extraction from texts or anaphora resolution). It is important to study the feasibility of integrating *GRASP*-it seamlessly in such systems. In these applied systems, the requirements for sub tasks often concern computational complexity and cost of execution. Therefore, this experiment is designed to study the gain/loss in performance relative to computing time in two different settings.

In order to maintain equivalence between the number of examples for each class, we do not consider cases of multiple classification in this evaluation. Regardless of the correctness of additional (fortuitous/incidental) predictions, these NPs were not anticipated as belonging to that (extra) class and are therefore not counted in the number of instances. Moreover, the number of these cases of multi-classifications is not predictable, nor evenly distributed across the types we consider.

4.4. Results

In the following, given a type of relation, we consider precision (P) of a class as the proportion of examples for which the class is correctly predicted relative to all instances predicted as belonging to that class. Recall (R) represents the ratio of examples for which the class is correctly predicted to all actual instances of that class.

4.4.1. Experiment 1 - Semantic Traits

An approach solely based on the semantics of A and B yields the results illustrated in Table 2. In order to assess separately the performance gain allowed by every trait included in the signature of terms, we report the results of 4 contrastive setting combinations of *GRASP*-it.

Setting	P (%)	R (%)	F1 (%)
H	67,3	65,9	65,3
H+SST	70,4	69,7	69,1
H+TRT	77,6	77	76,7
H+TRT+SST	**78**	**77,3**	**77,2**

Table 2: Average precision P (%) , recall R (%) , and F1 (%) scores achieved by the different settings of *GRASP*-it.

The components combined in these settings consist of the information stored while computing the signatures (thus, during the learning of rules): Hypernyms (*H*), Target for Relation Types (*TRT*), and Standard Semantic Types (*SST*).

Firstly, as a baseline, Hypernyms being lexical-semantic traits lead to an average *F1* score of $65,3\%$ which we deem satisfactory considering the potential sparsity of certain terms' hypernyms (e.g. we draw attention to terms A of the types *Characterisation (Ca), Agent (A) and Patient (P)* being all abstract entities for which it is delicate to identify un lexical hypernym). Secondly, we observe that both *SST* and *TRT* traits lead to notable improvements, the non-linear gain tends naturally to become smaller as the scores improve. *H* and *TRT* traits together on one hand and the *SST* on the other, are complementary as they each address specific description needs. *SST* provides the covering of standard typologies (e.g. providing the *_INFOSEM-THING-ABSTRACT* type that helps in the scenario discussed above), while *TRTs* and *H* (being terminological entries of lexical nature) bring forth a finer granularity made possible by an abundance in terminology.

With an F1 score of $77,2\%$, the most favourable setting is the one combining all semantic traits. Table 3 reports the results of the *H+TRT+SST* setting for every considered semantic type.

Type	P	R	F1
Origin (O)	100	86	92
Social Link (LS)	83	100	91
Holonymy (H)	78	86	82
Quantification (Q)	82	80	81
Agent (A)	71	93	81
Depiction (D)	88	73	80
Material (M)	78	83	80
Instrument (I)	77	80	78
Location (L)	84	70	76
Topic (T)	68	86	76
Patient (P)	74	76	75
Producer (AC)	76	73	74
Consequence (Co)	76	63	69
Possession (Po)	65	63	64
Characterization (Ca)	71	50	59
Average	**78**	**77,3**	**77,2**

Table 3: Percentages (%) of Precision (P), recall (R), and F1 scores achieved by the best semantic setting of our system (*GRASP-it*$_{(H+TRT+SST)}$), for every considered type of semantic relations.

We observe generally high results, though disparities exist depending on the type of relation being identified.

Specifically, the low recall for the *Characterization* (*Ca*) relation can be attributed to its limited representation in the database (approximately $10,000$ relations compared to the *Holonymy* (H) relation, which has over $10\ million$ relations), resulting in a low proportion of correctly annotated learning examples. The same applies to test examples (up to half of the cases). Additionally, in the case of

(*Ca*), generics are also poorly represented and often associated with a sense of the term that is not a property (see Section 4.5). It is noteworthy that there is maximum recall for the *social link* type (*LS*) carried by *relational* nominal heads, which allows for a lexical interpretation of this type and is well-synthesized by the created constraints. The Origin type (*O*) is precise due to its small set of general rules (a large number of rules could be merged), but it fails for examples that are not well-represented in the corpus. This case is interesting because instances of the Origin relation type are nearly nonexistent in *JDM* (29 relations), with Hypernyms facilitating the synthesis of effective rules.

In many respects, the followed protocol seeks to assess the model rigorously; indeed, the scores should be interpreted as a baseline to be improved through various treatments of classification processes using *GRASP*-it rules. It is worth noting the absence of morpho-syntactic heuristics (extra-semantic). Additionally, instances that were deemed erroneous (according to the corpus) but are actually valid for the predicted type are counted as errors. A multi-label evaluation would likely elevate the scores for each type (for example, the F1 score of the least well-classified type (Ca) would increase to 76%), however such an evaluation would require manual annotation.

4.4.2. Experiment 2 - Definiteness

Setting	**P**	**R**	**F1**
H+TRT+SST	78	77,3	77,2
H+TRT+SST+DEF	**80,3**	**79,8**	**79,5**
Origin (O)	100	90	95
Social Link (LS)	85	100	92
Holonymy (H)	81	100	90
Instrument (I)	88	80	84
Quantification (Q)	86	83	84
Material (M)	83	83	83
Depiction (D)	85	80	82
Location (L)	85	76	80
Agent (A)	67	96	79
Producer (AC)	79	76	77
Topic (T)	70	86	77
Patient (P)	72	70	71
Consequence (Co)	76	63	69
Possession (Po)	76	63	69
Characterization (Ca)	71	50	59

Table 4: Scores after inclusion of the definiteness trait (*GRASP-it*$_{(H+TRT+SST+\mathbf{DEF})}$), for every considered type of semantic relations compared to average scores in the solely semantic setting.

When considering definiteness, we include in the signature of the term B two distinct symbols corresponding to the presence (*resp.* absence) of a definite or indefinite determiner (*Det*, *NoDet*, *Def*, *NoDef*). One special case concerns named entities where none-withstanding a *NoDet* the *Def* attribute is "*forced*". Some examples : *chat du rabbin* => *Det* + *Def: - écran de cinéma* => *NoDet* + *NoDef* - *tableau de Chagall* => *NoDet* + *Def*.

As seen in Table 4, taking into account the definiteness trait improves the overall F1 score (compared to the exclusively semantic setting). Nevertheless, we observe some variability in improvement, and in two particular cases (agent and patient) a decrease in performance. The reason for this is that the trait of definiteness isn't typical of one unique type, rather of a subset of types. Its inclusion helps when deciding between two types for which the definiteness trait is decisive. On the contrary, it brings some confusion between types within the same subset of rules (for which the definiteness is not decisive).

4.4.3. Experiment 3 - Rules Trimming

In table 5, we highlight the differences in terms of the number of rules applied ($\#$**r**) and execution times (T) for the *Trim* and *No Trim* settings, which correspond respectively to a complete set of rule and a reduced one. The reduced rule set contains only rules that have not been used for a fusion operation. (T) is the duration for 450 test instances.

Setting	**P**	**R**	**F1**	$\#$**r**	**T (s)**
Trim	79,6	77,6	77	49	*25.42*
No Trim	80,36	80	79,8	1384	*92.78*

Table 5: Effects of the reduction of rules on performance scores and total execution times.

The execution time depends on the number of rules and on the size of the signatures. The more fusion occurs, the longer the signatures are. The trim setting leads to a tremendous improvement in computing time (a bit less than 4-fold) with a slight decrease in quality. This means that, as expected, systems that require responses in particularily low execution times could benefit from trimming the set of rules without suffering a significant degradation of the overall performance.

4.5. Analysis of Failure Cases

As a reminder, this work also serves as a tool to control the content of the knowledge base, and helping consolidate appropriate types of relations. The potential gaps in knowledge are highlighted through the analysis of the algorithm's *failure of classification cases*. Among the failure cases, approximately 75% of occurrences are directly attributed to the polysemy of term A and/or term B (combined). For instance, term A in "richesse du royaume" is interpreted in the sense of an *object/artifact* rather than the property of *wealth*, leading the system to classify it as a *location* (an object can be found in a place). Excluding cases of multiple classes (such as "*ombre d'un arbre*" or "*travail du réalisateur*", classified respectively as *(Ca)* and *(AC)*, and predicted by the system as *(D)* and *(A)* (which is also correct). The reasons of remaining errors include knowledge deficiencies (stemming from gaps in JDM) as seen in "*restaurant de cuisine végétalienne*" (where term B was unknown). The use of multiple prepositions is the cause of some failures ("*restaurant de fruits de mer*") due to misidentification of terms A and B (which is not always trivial). Furthermore, a few skill deficiencies (stemming from lack of processing) account for a small number of cases. Specifically, the management of semantic attribute dispersion in *JDM* across various morphological variants of a term: it happens, for intentional/valid reasons or due to the lack of knowledge, that some relations have not been propagated to all derivations of a given term. Regarding our system, this deficiency can be observed particularly at the level of the standard semantic type (SST). In the case of "*liste de films*" (*list of films*) supposed to be classified under the Quantification (Q) type, the standard types of the lemma "film" were missing in its plural variant. Although this failure case proved useful in indicating the state of the database concerning this type of relation, a normalization phase (specifically, a lemmatisation phase) would have circumvented this dispersion case.

5. Conclusion

We have presented the results of a study on genitive noun phrases in the form "A de B". The contributions of this work involved defining a non-exhaustive typology of semantic relations between terms A and B, producing a small corpus of examples annotated by these types, and proposing a learning algorithm for classification rules in the form of aggregated and semantically ordered constraint pairs. The goal is to achieve automatic classification of types that can be used in semantic text analysis and semantic lexicon consolidation.

The insights gained from this project span theoretical and applied domains. Firstly, we recognize the challenge posed by the possibility of including examples in multiple different types. Furthermore, although type identification heavily relies on semantics, it is not the sole criterion for classification decision-making. This implies that a high-performing algorithm should include a series of treatments, including data preparation and analysis (syntactic analysis, compound word identification, polysemy management, etc.).

The adopted symbolic approach has the advantage of being easily explainable. However, its results depend on the types considered, the quality of the data (balance of *weak/strong signals*, good representation of cases), and the richness and quality of the world knowledge base used.

Finally, the fact that rules are not mutually exclusive (and, in some cases, almost identical) presents a challenge that could not be resolved *a priori* – even for a human speaker – without sufficient context, such as in the examples: "*présentation de l'élève - portrait de Van Gogh - travail du ciment*" (*student presentation - Van Gogh portrait - cement work*) (*resp. could be one of : (A)/(P) - (A)/(D) - (A)/(P)*).

As with any corpus construction, the issue of distribution balance and coverage of traits/patterns in the learning set arises. This issue is particularly pronounced for a relatively small dataset. The question of coverage is challenging: intuitively, one might think that increasing the number of examples is necessary. However, the coverage of relation types in genitive forms follows a power law, meaning there are numerous specific cases in the long tail. These cases are difficult to quantify as they may correspond to figurative forms and are often non-prototypical examples (quasi-hapaxes) from a signature perspective. Additionally, prototypical forms corresponding to the beginning of the power law distribution are often abundant (for example, an impressive number of "<animal> of <place>"). Increasing the number of training examples with nearly identical signatures (represented by the same pattern model) has little long-term effect on learning quality. Furthermore, given the computational cost constraint posed by the need to integrate the solution into applied systems, it would be counterproductive to add examples related to an already known pattern. However, one could consider a strategy that exploits a fusion criterion (whether a rule is fusible or not) through incremental learning.

On the other hand, regarding test examples, increasing the number of instances may prove to be useful. We can expect, for a larger test set, a more or less significant decrease in performance as uncertainty increases, but this would allow for a better assessment and thus provide a more accurate idea of the system's quality. In the same vein, it would be useful to establish an evaluation that would assess the real impact of the *GRASP*-it results on applied NLP systems. Looking ahead, potential extensions of this work might include enriching the corpus through the application of our algorithm. A larger collection would allow a thorough evaluation and enable the utilization of various methods, particularly those requiring very large numbers of instances, such as neural learning approaches.

Distribution

Both the data collection and the algorithm are available at (Guenoune and Lafourcade, 2024). For demonstration and experimentation purposes, a webpage hosting a sample implementation of *GRASP*-it is available. A number of settings is proposed, namely the inclusion or not of the traits (*H, TRT, SST and DEF*), the trimming of rules and the setting of the fusion threshold are also possible. Both the training and the test can be done either with the proposed corpus or with another collection. Furthermore, it is not necessary for the semantic types to be restricted to those defined in this study (more or less types can be defined and used within the webpage). However, since the signatures built for this implementation are based on our knowledge base and its structure (the *JDM* structure), we should note that if another data collection is used, it is necessary to ensure that the examples comply with the requirements discussed in the *data preparation* section (4.1).

6. Bibliographical References

Chris Barker, Claudia Maienborn, Klaus von Heusinger, and Paul Portner. 2019. Possessives and relational nouns. *Semantics-noun phrases and verb phrases*, pages 177–203.

Asma Ben Abacha. 2012. *Recherche de réponses précises à des questions médicales: le système de questions-réponses MEANS*. Ph.D. thesis, Paris 11.

Jos De Bruin and Remko Scha. 1988. The interpretation of relational nouns. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Hani Guenoune. 2022. *Résolution des anaphores dans la communication électronique médiée: heuristiques et apport d'informations de sens commun*. Ph.D. thesis, Université de Montpellier.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramon Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, et al. 2020. Leveraging abstract meaning representation for knowledge base question answering. *arXiv preprint arXiv:2012.01707*.

Mathieu Lafourcade. 2007. Making people play for lexical acquisition with the jeuxdemots prototype. In *SNLP'07: 7th international symposium on natural language processing*, page 7.

Mathieu Lafourcade and Nathalie Le Brun. 2023. Apport du jeu pour la construction de connaissances: le projet jeuxdemots. *Technologie et innovation*, 8(4).

Frédéric Landragin. 2018. Étude de la référence et de la coréférence: rôle des petits corpus et observations à partir du corpus mc4. *Corpus*, 2(18).

Sebastian Löbner. 2011. Concept types and determination. *Journal of semantics*, 28(3):279–333.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Fifth international workshop on computational semantics (IWCS-5)*, pages 285–301.

7. Language Resource References

Guenoune, Hani and Lafourcade, Mathieu. 2024. *GenitiveNPS.txt.zip*. repository hosted at jeuxdemots.org. PID https://www.jeuxdemots.org/rezo-GEN1.php.

Lafourcade, Mathieu. 2024. *20240211-LEXICALNET-JEUXDEMOTS-FR-NOHTML.txt.zip*. repository hosted at Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier – LIRMM. PID https://www.jeuxdemots.org/jdm-about.php.

How Human-Like are Word Associations in Generative Models? An Experiment in Slovene

Mojca Brglez*, **Špela Vintar***, **Aleš Žagar**†
*Faculty of Arts, University of Ljubljana
Aškerčeva 2, Ljubljana
mojca.brglez, spela.vintar@ff.uni-lj.si
†Faculty of Computer and Information Science, University of Ljubljana
Večna pot 109, Ljubljana
ales.zagar@fri.uni-lj.si

Abstract

Large language models (LLMs) show extraordinary performance in a broad range of cognitive tasks, yet their capability to reproduce human semantic similarity judgements remains disputed. We report an experiment in which we fine-tune two LLMs for Slovene, a monolingual SloT5 and a multilingual mT5, as well as an mT5 for English, to generate word associations. The models are fine-tuned on human word association norms created within the Small World of Words project, which recently started to collect data for Slovene. Since our aim was to explore differences between human and model-generated outputs, the model parameters were minimally adjusted to fit the association task. We perform automatic evaluation using a set of methods to measure the overlap and ranking, and in addition a subset of human and model-generated responses were manually classified into four categories (meaning-, position- and form-based, and erratic). Results show that human-machine overlap is very small, but that the models produce a similar distribution of association categories as humans.

Keywords: word associations, generative models, T5, multilingual, Slovene

1. Introduction

Free word associations are a widely known technique of researching the human mental lexicon and have been used well before the emergence of psycholinguistics as a discipline (Galton, 1879). Having participants give oral or written associations to a cue word with as little reflection as possible may sound like a simple task, but as it turns out the responses given by different people show great variation both in the type of semantic relation governing the cue-response pair (cat -> dog vs. cat -> black vs. cat -> rat) and the individual association style (Fitzpatrick, 2007).

Since human associations to a large extent adhere to some patterns of semantic, syntactic or orthographic proximity, the emergence of vector-space meaning representations and early language models soon motivated a number of studies comparing different notions of relatedness in the human mental lexicon and that of a language model (see Section 2).

In this work, we describe an experiment which follows a similar aim, but for the first time such a comparison can be performed for Slovene, mostly because the human association dataset (SWOW-SL) for this language has been created only recently (see Section 3). We fine-tune two generative models to perform the task of responding to the prompt *"Which words do you associate with the word [WORD]?"* and select the parameters best suited to the association generation task. Since our aim is not to achieve maximum overlap between human and machine output but to better understand the workings of the artificial semantic space, we perform a series of evaluations. These include five different metrics to measure the concordance between the human and neural responses, and a qualitative evaluation through manual annotation in order to analyse the types of associations produced by humans and the language models.

In short, our key contributions in this work are:

- We construct and describe the first general dataset for Slovene word associations;
- We are the first to use generative models to explore word associations in language models;
- We are the first to test and evaluate the associations as output by generative models, using quantitative and qualitative methods in both English and Slovene.

2. Related Work

In recent years, several studies have attempted to evaluate the ability of vector space models to represent conceptual organization. Mandera et al. (2017) perform a detailed evaluation of correlations between human semantic spaces and corpus-based vector representations, whereby for the former they use semantic priming, semantic related-

The Workshop on Cognitive Aspects of the Lexicon, pages 42–48
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

ness judgements and word associations, while the vector-space models tested include count (LSA and HAL) and static neural models, referred to as predict models (skipgram and CBOW). They find that predict models, especially CBOW, consistently outperform traditional count models, and that the window size used in training plays a significant role in the performance of the models. They also report that a larger training corpus did not necessarily improve the results as in several experiments models trained on a smaller subtitle corpus outperformed those trained on UKWaC.

In an experiment by Nematzadeh et al. (2017) human word associations were compared to nearest neighbours suggested by word2Vec and GloVe, and they show that overall correlation is low and that static word embeddings fail to capture certain critical aspects of human associations.

The debate about common misconceptions about what word embeddings do or do not represent from a cognitive linguistics viewpoint was continued by Günther et al. (2019). One important emphasis, relevant also for our own experiments, is that while neural models are extremely powerful in producing quantitative representations of word meaning from (almost exclusively) textual data, the original idea behind Latent Semantic Analysis (LSA) was of it being not merely a computational but an explanatory tool, shedding light to how "word meanings are acquired through experience". For this reason, models trained directly on introspective data generally outperform corpus-trained ones (De Deyne et al., 2016).

Along similar lines, Jones et al. (2018) point out that association retrieval in humans is not symmetric, hence cosine distance may not be the best way to predict association strength. A more recent detailed discussion of the complexity of human associative behaviour and neural modelling is provided by Richie et al. (2022) who also train their GloVe model on English SWOW (De Deyne et al., 2018) and achieve good prediction results using a variety of asymmetric measures.

Although many studies have explored how human associations are represented in vector space models, they have all done so through indirect intrinsic measures (cosine similarity, probability distributions). To the best of our knowledge, our study is the first to directly explore the generation of associations with large language models. Secondly, the study is the first to examine the automatic generation of word associations for Slovene.

3. Datasets

3.1. English SWOW

For the training of the English model we use word association norms created in the English Small World of Words project (SWOW-EN, (De Deyne et al., 2018)). The data set consists of over 12,000 cue words and responses from over 90,000 participants, which makes it the largest resource of its kind for English. Responses were (and still are) collected through a web interface which presents association collection as a game, and participants were recruited via social media, e-mail and university websites. Over the years the SWOW project, originally developed for Dutch, grew into an ongoing world-wide study which currently covers 19 languages, including Slovene.

As a pre-processing step on SWOW-EN, we discarded the cue words which were labelled with a meaning gloss, as in *bat*, *bat(animal)*. The data was split into a training and testing set, whereby both subsets were sampled proportionally with regard to the PoS frequency.

3.2. Constructing SWOW-SL

The data collection for SWOW-SL[1] was supported by the generous help of Simon De Deyne (University of Melbourne) and Gert Storms (University of Leuven), the authors of the original SWOW project, who kindly added Slovene as another language on the SWOW platform, imported the data for the experiment and set up the localized web pages describing the task. The experiment for Slovene is the same as for other languages; a participant is consecutively shown 18 cue words and is asked to contribute up to 3 associations to each. At the end of the experiment the participant is shown some preliminary results, such as overlap with other participants and basic project statistics.

The selection of cue words for Slovene was based on the frequency lexicon from the Gigafida 2.0 corpus (Krek et al., 2020), whereby we limited the part-of-speech for cues to nouns, adjectives, verbs and adverbs, and then selected lemmas from the top 500-1500 frequency-ranked items, but removing proper names, acronyms and adjective-adverb duplicates (e.g. *dober* - *dobro*). The data collection for Slovene started in November 2023 and has reached 671 participants and the time of writing this article. Our dataset was constructed with responses up to January 10, 2024. To that point, each cue word has received responses from 8–10 different respondents, with approximately 17 unique responses per cue word on average.

[1] https://smallworldofwords.org/sl

Before dividing the Slovene dataset into testing and training splits, we apply lowercasing in order to normalize the responses. Because previous works (De Deyne and Storms, 2008) have shown that response categories can vary by the PoS of the cue, we sample the test (evaluation) set proportionately from each PoS according to its relative frequency.

Dataset	train	test
SWOW-en	10.946	611
SWOW-sl	949	51

Table 1: Datasets and data splits. The number refers to the number of cues.

Table 2 presents the structure of a training example for Slovene and English. Both consist of an immutable input prompt integrating the **cue** and the expected output with all the human *responses*.

4. Experimental Setup

4.1. Models

We employed two state-of-the-art models, namely SloT5 (Ulčar and Robnik-Šikonja, 2023) and mT5 (Xue et al., 2021). Both models are rooted in the transformer architecture, characterized by an encoder-decoder framework, and have been pre-trained to generate text effectively.

For our experimentation, we utilized preprocessed datasets described in the previous section. The SloT5 model was deployed in a monolingual setting, focusing solely on the Slovene language version of the dataset. In contrast, the mT5 model, known for its multilingual capabilities, was trained on a concatenated dataset comprising both Slovene and English versions, thereby facilitating a multilingual experiment.

4.2. Evaluation

To assess the trained generative models on how well they align with human associative networks, we evaluate 1) the overlap between human and model responses, 2) the ranking of responses, and 3) the categories of the responses. We employ five distinct automatic metrics for the first and second aspect, and perform manual annotation on a sample of data for the third aspect.

Automatic Metrics To assess the performance of our trained generative models, we employ five distinct automated metrics, including 4 similarity and 1 distance metric:

1. **Jaccard** similarity compares the size of the intersection of two sets to their union, i.e. it provides a measure of how much overlap exists between two sets of items with regard to the whole number of distinct items in both sets.
2. Rank-based Overlap (**RBO**, Webber et al., 2010), used especially in information retrieval, evaluates the similarity between two sets by considering both the overlap and the ranking of items. This metric assigns a larger weight to items appearing higher in the list of the gold standard (here, our the human responses).
3. **Precision** and **Recall**: Conventional metrics used to evaluate the accuracy of the responses.
4. Word Mover's Distance (**WMD**, Kusner et al., 2015) assesses the minimal "transportation cost" needed to move from one set of word embeddings to another. We use this metric to measure semantic similarity beyond direct word matches, where a lower score means a shorter distance travelled and thus a more similar set of words.

To enhance the clarity and interpretability of the automatic metrics, we also assess each metric on the overlap of human-generated associations. We simulate the human subsets by first concatenating all responses for a specific cue according to their associative strength. Then, we distribute these responses into two groups, alternating between each group for every response in the ranked list.

Manual Annotation of Response Categories In addition to evaluating the performance of the trained models with the methods described above, our aim was to better understand the behaviour of neural models with respect to the types of responses they produce. As proposed by (Fitzpatrick, 2007), human associations can be classified into four broad categories, namely:

- **Meaning-based association**: *x* means the same as *y*, *x* and *y* come from the same lexical set, *x* and *y* have some other conceptual link
- **Position-based association**: *y* follows or precedes *x* directly or with words between them
- **Form-based association**: *y* is *x* plus or minus affix, *y* looks or sounds similar to *x*
- **Erratic association**: *y* has no decipherable link to *x*

We first performed a manual annotation of categories on a random selection of 50 Slovene cues with human responses. A total of 996 responses were classified into the above categories by at least two annotators, followed by a curation phase

Input **Output**	Na kaj pomisliš ob besedi **umetnik**? [What comes to mind when thinking about the word **artist**?] Najprej pomislim na besede: *jaz, galerija, flamenko, slika, slikar, poet, slika, kri, beretka, lev, molitev, oblaki, dolina šentflorijanska, kreator, barve, pavarotti, len, ljubezen, radost, poslanstvo, čustva, picasso, lepo* [I first think of the words: *me, gallery, flamenco, painting, painter, poet, painting, blood, beret, lion, prayer, clouds, St. Florian valley, creator, colours, pavarotti, lazy, love, joyfulness, calling, emotions, picasso, beautiful*]
Input **Output**	Which words do you associate with the word **arctic**? I associate this word with: *cold, ice, snow, polar bear, circle, penguin, polar, north, white, tundra, Antarctic, North Pole, freezing, icy, ocean, penguins, Antarctica, bear, fox, freeze, monkeys, polar bears, pole, frozen, glacier, iceberg, igloo, roll, air, Arctic circle, Arctic ocean, Aurora, bears, blue, char, chilly, clear, collapse, conditions, continent, cruise ship, dappled, enema, Eskimo, expedition, explorer, extreme, far, far away, flexible, free, frigid, frost, hare, icebreaker, lights, monkey, orca, owl, p, pudding, region, resonance, Russia, sadness, Santa, seals, ship, slicer, snot, snowy, software, spare, spontaneous, temperature, war, wind, winter, wolf, zone*

Table 2: A Slovene and an English training example, consisting of an input prompt integrating the cue (in bold) and the expected output with all the human responses (in italic)

.

to resolve inter-annotator disagreements. The inter-annotator agreement between pairs of annotators was on average moderate with Cohen's Kappa score of 0.507, but the values varied greatly amongst pairs ranging from meager 0.147 to 0.80. Since the annotators were students who were given only a brief training before the actual annotation, many inconsistencies were resolved later through discussion and curation. On the other hand the task itself is somewhat ambiguous as many responses could legitimately be assigned several categories.

The category frequencies of human responses by cue part-of-speech are given in Table 3. Over one half of responses fall into the meaning-based category, with verbal cues deviating from the typical distribution of categories by favouring position-based associations. It would appear that verbs as cues are stronger triggers for collocational patterns than other part-of-speech.

A similar classification of responses was then performed for a randomly selected set of 10 cue words for all three evaluated models: SloT5, mT5-SL and mT5-EN.

PoS	Erratic	Form	Meaning	Position
Adj	12	13	72	60
N	52	27	311	114
Adv	19	2	68	27
V	16	15	71	117
Total	99	57	522	318

Table 3: Categories in human annotated associations by PoS of the cue word

5. Results

To fine-tune the SloT5 and mT5 models for our specific research task, we employed a consistent set of hyperparameters across both SloT5 and mT5 models. These included a learning rate of 5×10^{-5}, a training span of 10 epochs, a batch size of 8, and the $AdamW$ (Loshchilov and Hutter, 2019) optimizer.

For the inference phase, careful consideration was given to the selection of parameters with the aim of preserving the model parameters (i.e. the existing network and representations) at their default values and adjusting them only slightly to obtain structurally sound outputs and to reduce repetitive behaviour from the models. The parameters configured were as follows: sampling was $enabled$ to introduce variation in the outputs, the maximum sequence length was set to 128 tokens, the top-k sampling was $disabled$ to prevent constraining the sampling space, a repetition penalty of 1.2 was applied to diminish redundancy in the text generation, and the nucleus sampling threshold was established at 0.8 to manage the diversity of the generated content.

5.1. Evaluation

Automatic Metrics The automatic evaluation of results shows extremely low overlap between the human word associations and the model-generated word associations. As shown in Table 4, the overlap, ranking and semantic similarity of responses is much higher for human subsets than for any of the trained generative models. Note that the deviations are much higher for the Slovene human-human subsets due to the small size of the dataset. Overall, the multilingual model performs marginally better on the English than on the Slovene dataset according to Jaccard, Precision, and WMD metrics. On the other hand, between the monolingual and multilingual T5 model for Slovene, the monolingual performs much better, achieving a higher score on all the five metrics.

	model	RBO	Jaccard	Precision	Recall	WMD
SL	human	0.22 ± 0.17	0.15 ± 0.1	0.24 ± 0.15	0.24 ± 0.15	0.76 ± 0.15
	sloT5	**0.05** ± 0.06	0.03 ± 0.03	0.05 ± 0.04	**0.09** ± 0.07	**0.95** ± 0.06
	mT5	0.02 ± 0.08	0.01 ± 0.02	0.03 ± 0.06	0.02 ± 0.03	1.05 ± 0.05
EN	human	0.43 ± 0.08	0.3 ± 0.05	0.46 ± 0.06	0.46 ± 0.06	0.3 ± 0.07
	mT5	0.03 ± 0.04	**0.04** ± 0.02	**0.13** ± 0.08	0.05 ± 0.03	**0.95** ± 0.05

Table 4: Results of automatic metrics for word associations overlap and their standard deviations. In bold: overall best score by trained models, underlined: best score on the Slovene dataset. Note that for WMD, which is a distance measure, a lower score is better.

Manual Evaluation Manual annotation of responses (see Table 5) produced by the models first revealed a much higher ratio of erratic responses (human 0.11 vs. 0.47-0.61 in models). Erratic responses are those where no meaningful relation or connection between the cue and response can be found, and such responses are typically rare in humans. Conversely, around a half of the models' associations are relatively far-fetched and thus labelled as erratic, and we can associate the ratio of erratic responses with the overall performance of the model rendering SloT5 as the winner amongst the three.

- SloT5: napisati [write] -> pes [dog], še [still], pisk [whistle], stranica [edge], informatika [informatics], ...
- mT5-SL: napisati [write] -> pravijo [they say], slikovati [*unknown word*], dogovoriti [arrange], požičiti [*unknown word*], obrazi [faces], ...
- mT5-EN: ecological -> white, creole, furry, cheerful, lively, instinctive, dark, ...

For the other three categories, Form-based, Meaning-based and Position-based respectively, the distribution of the models' responses is surprisingly similar to human categories, with very few form-based responses and a good measure of meaning-based ones (see Figure 1).

Another observation concerning the output of the SloT5 and mT5-SL models is the occurrence of unknown (invented) words as responses to cues. While the SloT5 model hardly ever forms an unexisting word (0.01%), in the mT5 model every 5th response is a newly created word (e.g. *ključak, obkreče, kuzko, nadvid, slikovati, požičiti, snemeti, avtores, pohnost, ...*

6. Discussion

Since the purpose of our experiment was to compare the associative behaviour of fine-tuned mono- and multilingual models with human association norms, low overlap between them does not necessarily mean failure. Thus, we did not use the now-popular instruction-based large language models

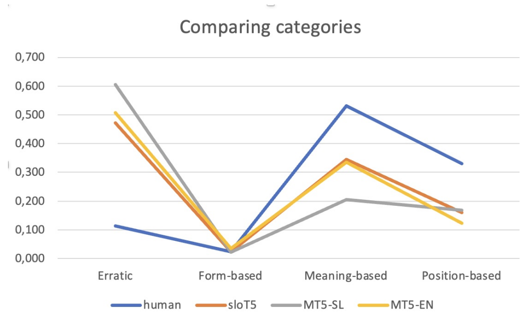


Figure 1: Distribution of categories in human responses and the three models

model	Erratic	Form	Meaning	Position
human	0.11	0.02	0.53	0.33
SloT5	0.47	0.02	0.34	0.16
mT5-SL	0.61	0.02	0.20	0.17
mT5-EN	0.51	0.03	0.21	0.12

Table 5: Distribution of categories in human responses vs. trained models

because they do not align with our experiment goal. In our experiment, we deliberately preserved model parameters at default values and see the results obtained as a kind of a baseline for a new language, for which no such study had been performed before.

The results for word associations overlap are generally low, but consistently best with the monolingual SloT5 model. Results also show that a much larger set of training data, which was the case for English, does not improve the alignment with human responses - in either the English or Slovene inference task.

The manual classification of human responses and predictions into categories shows that the models behave in a manner rather similar to humans in that they generate - if not erratic - mostly meaning-based associations which entail synonyms, hypo- and hypernyms and other more loosely related words (e.g. *natančno* [exact] -> *točno* [precise], *umetnost* [art] -> *igra* [play], *partner* [partner] -> *odnos* [relationship]). Similar to human norms, predicted words for verbal cues contain a slightly higher number of position-based associations (e.g. *napisati* [write] -> *odgovor* [reply], *knjigo* [book], *besedilo* [text]). We speculate that the fact that the multilingual mT5-SL generated a high number of non-existing words in the erratic category, compared to both SloT5 and mT5-EN, is due to a lower quality and quantity of pre-training data.

Our research is limited in that it uses a rather small dataset for Slovene, where the number of human responses collected for each cue is considerably lower than for English. Later versions of the dataset may prove better in this respect. Another limitation is that tha manual annotation comprised only a relatively small random sample of responses, so that the overall distribution might be different for a more representative sample. We also assume that results would be different when employing newer and larger language models.

7. Conclusion

The first contribution of our work is the creation of a new resource, the first version of SWOW-SL containing human associations to 1000 Slovene cues contributed by over 600 participants, and created under the auspices of the "parent" Small World of Words project (De Deyne et al., 2018). We then use this dataset to fine-tune a monolingual T5 and a multilingual mT5 model (as well as an English one for comparison) for the word association task, but without attempting to optimize the parameters. The predictions of the models are evaluated using 4 automatic metrics, namely Jaccard, rank-biased overlap, precision and recall and Word Mover's Distance. Results show that the overlap between human and model-generated responses is very low, and that the better model for Slovene is the monolingual one. A manual classification of responses into categories is performed in order to better understand the behaviour of the models. While all models generate a high number of erratic responses (between 47 and 61 percent), the distribution of meaningful responses amongst the meaning-based, position-based and form-based categories closely resembles human norms.

8. Acknowledgements

The authors thank Simon De Deyne of Melbourne University for his kind help in setting up the SWOW experiment for Slovene, which makes possible the ongoing collection of association data for Slovene. This research was partly supported by the ARIS research programmes P6-0215 Slovene Language - Basic, Contrastive, and Applied Studies and P6-0411 Language Resources and Technologies for Slovene.

9. Bibliographical References

Álvaro Cabana, Camila Zugarramurdi, Juan Carlos Valle-Lisboa, and Simon De Deyne. 2023. The "small world of words" free association norms for rioplatense spanish. *Behavior Research Methods*, pages 1 – 18.

Simon De Deyne, Danielle Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 1861–1870.

Simon De Deyne and Gert Storms. 2008. Word associations: Network and semantic properties. *Behavior research methods*, 40(1):213–231.

Tess Fitzpatrick. 2007. Word association patterns: unpacking the assumptions. *International Journal of Applied Linguistics*, 17(3):319–331.

Francis Galton. 1879. Psychometric experiments. *Brain*, 2(2):149–162.

Fritz Günther, Luca Rinaldi, and Marco Marelli. 2019. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, 14:1006 – 1033.

Michael N. Jones, Thomas M. Gruenenfelder, and Gabriel Recchia. 2018. In defense of spatial models of semantic representation. *New Ideas in Psychology*, 50:54–60.

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. 2020. Gigafida 2.0: the reference corpus of written standard slovene. In *Proceedings of the twelfth language resources and evaluation conference*, pages 3340–3345.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.

Ken McRae, George Cree, Mark Seidenberg, and Chris Mcnorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37:547–59.

Douglas Nelson, Cathy Mcevoy, and Simon Dennis. 2012. What is free association and what does it measure? *Memory and Cognition*, 28:887–899.

Aida Nematzadeh, Stephan C Meylan, and Thomas L Griffiths. 2017. Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *CogSci*.

Russell Richie, Ada Aka, and Sudeep Bhatia. 2022. Free association in a neural network. *Psychological Review*.

Paul Rozin, Nicole Kurzer, and Adam B. Cohen. 2002. Free associations to "food:" the effects of gender, generation, and culture. *Journal of Research in Personality*, 36(5):419–441.

Avijit Thawani, Biplav Srivastava, and Anil Singh. 2019. SWOW-8500: Word association task for intrinsic evaluation of word embeddings. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 43–51, Minneapolis, USA. Association for Computational Linguistics.

M Ulčar and M Robnik-Šikonja. 2023. Sequence-to-sequence pretraining for a less-resourced slovenian language. *Frontiers in Artificial Intelligence*, 6:932519–932519.

Ivan Vulić, Douwe Kiela, and Anna Korhonen. 2017. Evaluation by association: A systematic study of quantitative word association evaluation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 163–175, Valencia, Spain. Association for Computational Linguistics.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28:20.

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.

Peiran Yao, Tobias Renwick, and Denilson Barbosa. 2022. WordTies: Measuring word associations in language models via constrained sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5959–5970, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Špela Vintar and Amanda Saksida. 2023. The anatomy of specialized knowledge: Comparing experts and non-experts through associations, frames and language models. *Lexicographica*, 39(1):165–190.

10. Language Resource References

Small World of Words - downloadable resources, `https://smallworldofwords.org/en/project/research`

Idiom Complexity in Apple-Pie Order: the Disentanglement of Decomposability and Transparency

Irene Pagliai
University of Göttingen
irene.pagliai@uni-goettingen.de

Abstract

Both decomposability and transparency investigate the interplay between literality and figurativity in idioms. For this reason, they have often been merged. This study argues that idiom decomposability and transparency are related but conceptually different constructs, thus advocating for their distinction. Leveraging a normed lexicon of Italian and English idioms, the respective effects of decomposability and transparency on idiom meaning recognition are explored via statistical modeling. Results show the two variables contribute differently to idiom meaning recognition in the two languages, while the absence of collinearity underscores their distinct contributions. Based on this empirical evidence, the study finally proposes FrameNet and MetaNet as computational tools for modeling idiom decomposability and transparency. This study thus not only substantiates the separation of idiom decomposability and transparency, but also sets a foundation for future interdisciplinary research to bridge the gap in idiom research between empirical psycholinguistics, cognitive linguistics and computational applications.

Keywords: idioms, decomposability, transparency

1. Introduction

Idioms are multiword expressions bearing a figurative meaning (Cacciari and Tabossi, 2014; Wagner, 2021). The typical interpretation of the idiom *spill the beans* does not indicate the actual spilling of a can of beans, but rather the revelation of a secret. Idioms' distinctive feature is therefore their duality between a compositionally derivable literal meaning and a conventionally associated figurative meaning (Sprenger et al., 2006).

Importantly, idioms constitute a heterogeneous category, varying across several dimensions (Langlotz, 2006; Wulff, 2013). Notably, an idiom's degree of decomposability and transparency are essential for an in-depth analysis of the interplay between an idiom's literality and figurativity (Geeraerts, 2003, 1995; Carrol et al., 2018).

Decomposability refers to the extent to which the figurative meaning of an idiom can be broken down by linking its figurative semantic components to its literal syntactic elements (Sailer, 2021). For instance, with *spill the beans*, the action of spilling corresponds to revealing, and the beans can be mapped onto the secrets being disclosed (Nunberg et al., 1994). Thus, decomposability focuses on the interface between an idiom's syntax and semantics. This is clearly outlined in Geeraerts (2003), who proposes the term "isomorphism", i.e., "a one-to-one correspondence between the formal structure of the expression and the structure of its semantic interpretation" (p. 438).

Transparency refers to the possibility of establishing a synchronic relationship between an idiom's literal and figurative meanings in their entirety (Sailer, 2021; Moreno, 2005). This "semantic bridge" (Dobrovol'skij, 2016, p. 23) works as the rationale for how a multiword expression has been semantically extended from its literal meaning to its idiomatic interpretation, and is grounded in cognitive-conceptual mechanisms like metaphors, metonymies, and encyclopedic knowledge (Kovecses and Szabco, 1996; Carrol et al., 2018). In the case of *spill the beans*, the sudden visibility of the spilled beans metaphorically mirrors the revelation of the secret, which, again metaphorically, has come out of its container and is therefore no longer under control.

Since both decomposability and transparency examine idioms' literal-figurative interplay, prior research on idiom features has often conflated them together within a single variable (see the discussions in Hubers et al., 2019; Michl, 2019; Carrol et al., 2018). While agreeing that the two variables are correlated (Carrol et al., 2018), we claim that they are different in kind (Geeraerts, 2003, 1995; Hubers et al., 2019; Carrol et al., 2018).

In addition, the distinction between decomposability and transparency aligns well with the hybrid model of idioms (Cutting and Bock, 1997; Cacciari and Tabossi, 1988; Sprenger et al., 2006; Libben and Titone, 2008; Titone and Connine, 1999). This account claims that idioms are encoded in the mental lexicon in a hybrid manner, through a multi-level interconnection of literal and figurative components. In this regard, Sprenger et al. (2006) argues that "idioms are both unitary and compositional, although at different levels of their cognitive representation" (p. 174). Differentiating between decomposability and transparency provides precise analytical tools with which to investigate with finer granularity the inherently dual nature of idioms. On the one hand, decomposability focuses on the syntax-semantics

The Workshop on Cognitive Aspects of the Lexicon, pages 49–55
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

interface; on the other hand, transparency targets the connections between the semantic and conceptual levels.

The aim of the present work is precisely to provide empirical evidence in support of the distinction between idiom decomposability and transparency. This is done by conducting an exploratory analysis on the respective effects of decomposability and transparency on idiom meaning recognition in two languages, Italian and English. Idiom meaning recognition is operationalized through the variable *objective knowledge*, which describes the correct identification of an idiom's figurative meaning by a speaker (see also Hubers et al., 2019). To verify the effects of decomposability and transparency on idiom objective knowledge, a cross-linguistic normed idiom lexicon where the variables have been quantitatively assessed ("normed") by native speakers (Pagliai, 2023) is employed.

Finally, one additional purpose is to foster interdisciplinary research on idiom variables and modeling. The analysis conducted here is psycholinguistic in nature, and is intended to be exploited as an empirical basis to support further research in cognitive linguistics, thus responding to the advocacy of Espinal and Mateu (2010) for a cognitive (psycho)linguistic approach. For this reason, the final part of the paper illustrates how idiom decomposability and transparency can be effectively modeled through two computational resources based on cognitive linguistics theories: FrameNet (Ruppenhofer et al., 2006) and MetaNet (Petruck, 2016).

2. Methods

The cross-linguistic lexicon comprises 150 pairs of Italian and English idioms sharing similar meanings (Pagliai, 2023). The dataset was obtained through the implementation of a cross-linguistic norming study in which idioms were normed by native Italian and English-speaking participants for a number of variables: familiarity, meaningfulness, objective knowledge, literal plausibility, decomposability, and transparency. All variables were operationalized on a 1 to 5 Likert scale, with the exception of objective knowledge. This was presented as a dropdown option including three idiom paraphrases from which to select the correct one. In the present analysis, objective knowledge provides the required measurement of idiom meaning recognition (for more details regarding the dataset creation, as well as the variables' choice, definition, and operationalization, please refer to Pagliai, 2023).

To assess the distinct impacts of decomposability and transparency on idiom objective knowledge in the two languages, generalized linear mixed models (GLMMs) were fitted using the `lme4` package (Bates et al., 2015) in R (v4.3.2, R Core Team, 2023). Maximum cross-linguistic comparability was ensured by employing identical model structures for both languages. Objective knowledge was set as dependent variable, with decomposability and transparency serving as predictors. To account for its influence, meaningfulness (the subjective degree of confidence of knowing the meaning of an idiom) was also included as an additional predictor[1]. All predictors were centered; the dependent variable objective knowledge was encoded using treatment coding, with "wrong" responses coded as 0 and "correct" responses as 1 . Each model incorporated random intercepts for both participants and items. Due to issues with models' convergence, no random slopes were included (Barr et al., 2013).

To ensure that each predictor variable uniquely contributed to idiom meaning recognition, multicollinearity was tested. Variance inflation factors (VIFs) for both Italian and English models were calculated using the package `performance` (Lüdecke et al., 2021). The same package was exploited for conditional and marginal R^2 calculation.

3. Results

Model results are shown in Table 1. For Italian idioms, the estimated log-odds of objective knowledge significantly increase by 0.82 (*SE* = 0.08, *p* < .001) for each unit increase in meaningfulness, equivalent to an odds ratio (OR) of 2.27, indicating a substantial positive impact. Conversely, decomposability presents a non-significant negative effect (β = -0.04, *SE* = 0.08, *p* = 0.643; OR $\approx$ 0.96). Transparency shows a significant positive association, with a log-odds increase of 0.27 (*SE* = 0.09, *p* = 0.004), corresponding to an OR of approximately 1.31.

Turning to English idioms, the same positive effect for meaningfulness is found, with an identical log-odds increase of 0.82 (*SE* = 0.06, *p* < .001; OR $\approx$ 2.27). In contrast with Italian, decomposability exhibits a significant positive relationship (β = 0.21, *SE* = 0.08, *p* = 0.009; OR $\approx$ 1.23). The effect of transparency, while positive, does not reach significance (β = 0.12, *SE* = 0.08, *p* = 0.115; OR $\approx$ 1.13).

The different effects of decomposability and transparency on idiom objective knowledge across the two languages are illustrated in Figure 1 for decomposability and Figure 2 for transparency. The plots capture the impact of the two variables on the predicted probability of correctly guessing id-

[1]The addition of familiarity (the subjective frequency with which a speaker uses and hears/reads an idiom) as a predictor was also considered. However, comparisons based on the Akaike Information Criterion (AIC) revealed that models excluding familiarity performed better in both Italian and English.

Italian				English			
Fixed effects	β	***SE***	***p***	**Fixed effects**	β	***SE***	***p***
(Intercept)	4.57	0.25	<.001	(Intercept)	3.62	0.25	<.001
Mean	0.82	0.08	<.001	Mean	0.82	0.06	<.001
Deco	-0.04	0.08	0.643	Deco	0.21	0.08	0.009
Tra	0.27	0.09	0.004	Tra	0.12	0.08	0.115
Random effects	**Variance**	**Std. Dev.**		**Random effects**	**Variance**	**Std. Dev.**	
Intercept: items	1.52	1.23		Intercept: items	1.97	1.40	
Intercept: participants	0.32	0.56		Intercept: participants	0.35	0.59	
Cond. R^2 = 0.46; Marg. R^2 = 0.16				Cond. R^2 = 0.55; Marg. R^2 = 0.23			

Table 1: Comparative summary of GLMMs predicting idiom objective knowledge as a function of meaningfulness, decomposability, and transparency. Model outcomes for Italian on the left, for English on the right.

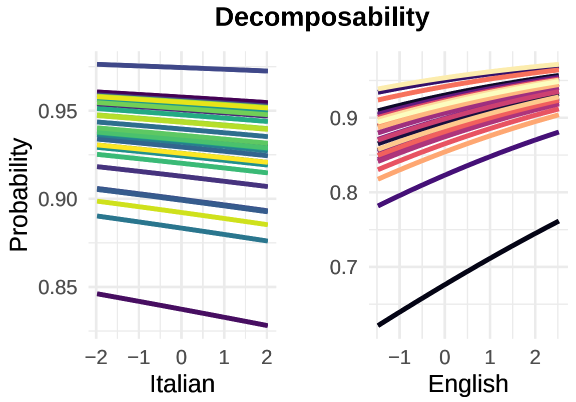


Figure 1: Effect of decomposability on the predicted probability of correctly identifying the meanings of Italian (left) and English (right) idioms, with by-participant variation (random intercepts).

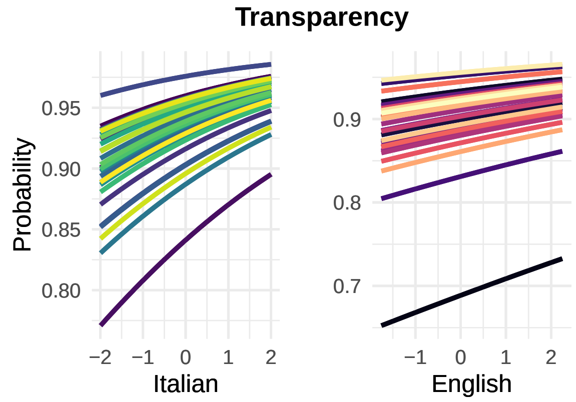


Figure 2: Effect of transparency on the predicted probability of correctly identifying the meanings of Italian (left) and English (right) idioms, with by-participant variation (random intercepts).

	Italian			English		
	Mean	**Deco**	**Tra**	**Mean**	**Deco**	**Tra**
VIF	1.05	1.19	1.19	1.02	1.41	1.43

Table 2: VIFs for meaningfulness, decomposability and transparency in Italian and English.

ioms' figurative meaning, while accounting for by-participant variation. For Italian, as decomposability increases, there is a slight downward trend in the probability of correct idiom knowledge. In contrast, for English, increased decomposability corresponds with a significant rise in the likelihood of idiom objective knowledge, as indicated by the slopes' upward trajectory. Transparency is associated with an increase in the probability of correct idiom meaning recognition in Italian, a trend that is present in English as well, although weaker and not supported by statistical significance.

Regarding collinearity investigation, results are reported in Table 2. All VIF values are below 1.5 in both Italian and English. This indicates absence of multicollinearity (Winter, 2019), confirming that each predictor variable—including decomposability and transparency—provides a unique and distinct contribution to idiom meaning recognition in both languages.

4. Discussion

The results underscore the different impact of decomposability and transparency in the two languages. For Italian speakers, the transparency of the relationship between literal and figurative meaning plays a key role in identifying the correct idiom paraphrase, while English speakers rely more on the isomorphism between syntactic and semantic structures. Moreover, the direction of the effect associated with decomposability differs cross-linguistically, being positive in English and negative in Italian.

This cross-linguistic difference could stem from the characteristics of the dataset, as Italian speakers showed higher knowledge of the idiom sample than English speakers, with respective overall objective knowledge scores of 96% and 89% (percentages obtained by calculating mean objective knowledge across all participants and idioms within each language). This difference might suggest that higher idiom knowledge leads to greater reliance on the conceptual link between literal and figurative meanings for idiom recognition. Conversely, with less knowledge, speakers may depend more on the literal-syntactic structure, possibly due to figurative meanings being less accessible in the mental lexicon. This would render grasping the semantic relationships between idioms' literal and figurative meanings more challenging.

Nevertheless, in the dataset, idiom knowledge is on average high across both languages. Indeed, when resorting to real idioms, it is not uncommon to find that speakers are highly familiar with them (Tabossi et al., 2011; Bulkes and Tanner, 2017), given idioms' recurrent presence in everyday language experiences (Jackendoff, 1995; Searle, 1975). This dataset feature, together with the fact that this is an exploratory study, calls for future confirmatory analysis. To ascertain whether the observed differences can be attributed to the level of idiom knowledge, follow-up research may involve ad-hoc created cross-linguistic idioms along with their paraphrases, to simultaneously control the level of idiom knowledge, decomposability, and transparency. With less skewed data, future studies could more reliably test the hypothesis that transparency's impact is more pronounced for idioms with which speakers are highly familiar, while decomposability plays a more significant role for less known idioms.

Collinearity was not detected in either Italian or English, reinforcing the view that decomposability and transparency are related but distinct variables. This cross-linguistic consistency strengthens the foundation for expanding research in cognitive linguistics relative to these two dimensions of idiomatic variation. Notably, the computational tools FrameNet (Ruppenhofer et al., 2006) and MetaNet (Petruck, 2016) could enhance our capacity to model decomposability and transparency effectively.

FrameNet is a lexical resource grounded in Frame Semantics theory (Fillmore, 2006). It represents word meanings through frames, "coherent schematizations of experience" (Fillmore, 1985, p. 223) describing situations, events, or objects. Frames are evoked by lexical units (form-meaning pairs; Cruse, 1986), and each frame includes a unique set of Frame Elements (FEs) to detail the roles and participants within these scenarios. MetaNet is a computational resource rooted in Conceptual Metaphor Theory (Lakoff and Johnson, 2008). It is designed to systematically capture and organize metaphors across languages, while highlighting the connections between concepts.

In FrameNet, frames are connected in a hierarchical network (Ruppenhofer et al., 2006, p. 73). Since transparency has been defined as a semantic relationship between the two idiom meanings, it can be modeled as a new literal-figurative frame-to-frame relation linking together the literal and the figurative frames evoked by an idiom. Building on this foundation, the MetaNet database, with its extensive repository of conceptual metaphors, can supply the metaphors that underpin the literal-figurative frame-to-frame relation. As for decomposability, it can be modeled as the possibility of establishing a mapping between the FEs involved in the two frames. Currently, FrameNet includes relationships between FEs, but only within the same frame (Ruppenhofer et al., 2006, p. 21). The decomposability of an idiom, conversely, depends on the possibility of establishing a relation between FEs belonging to two different frames: one literal and the other figurative.

Consider Figure 3, which visually exemplifies how to model the decomposability and transparency of the idiom *spill the beans* by resorting to the tools provided by FrameNet and MetaNet. In isolation, the idiom evokes two frames: the literal Cause_fluidic_motion ("An Agent or a Cause causes a Fluid to move") and the figurative Reveal_secret ("A Speaker reveals Information that was previously secret to an Addressee"). Focusing on the bottom of the figure, let us first consider decomposability. The noun phrase "the beans" corresponds to the FE FLUID ("the entity that changes location and moves in a fluidic way") in the literal frame Cause_fluidic_motion, while it corresponds to the FE INFORMATION ("the content that the Speaker reveals to the Addressee") in the figurative frame Reveal_secret. The mapping between the two FEs underlies the interpretation whereby the spilled beans (FLUID) correspond to the disclosed secrets (INFORMATION). Therefore, decomposability acts at the intersection of syntax and semantics, as evidenced by the triangular relationship connecting the nominal constituent "the beans" and the two FEs, one in the literal and the other in the figurative frame.

Moving upward, transparency can be conceptualized as the frame-to-frame relation that is established between the literal frame Cause_fluidic_motion and the figurative frame Reveal_secret in their entirety (as opposed to decomposability, which is a relationship between frame sub-components). Further enriching this analysis, MetaNet provides the underlying metaphors that scaffold the literal-figurative frame-to-frame relation. EXISTENCE IS VISIBILITY

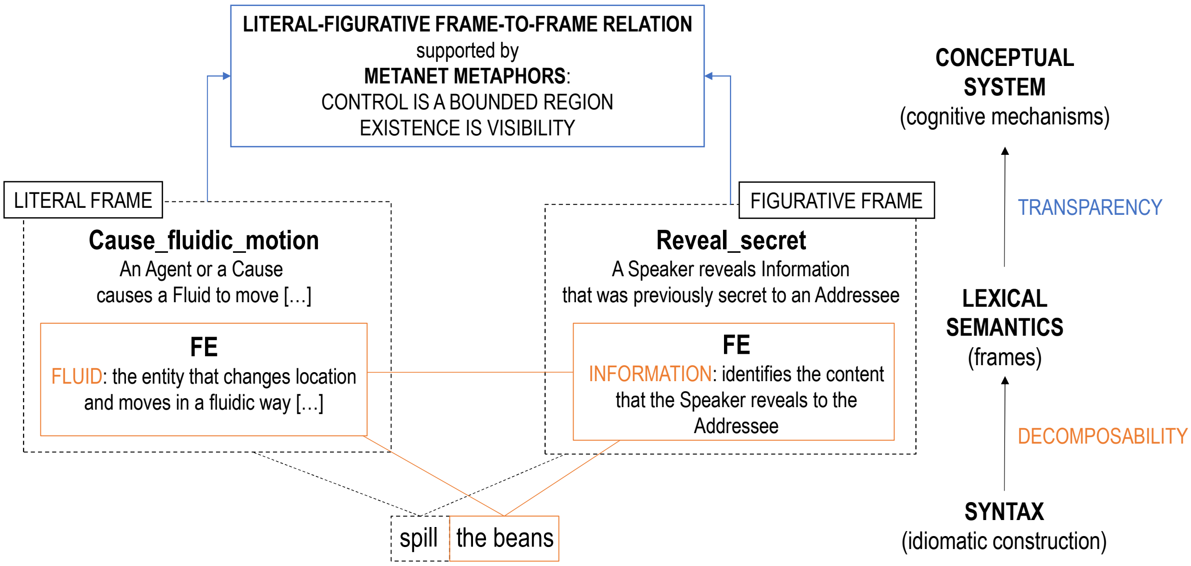


Figure 3: Modeling of the decomposability and transparency of *spill the beans*, resorting to the tools of FrameNet and MetaNet. On the right is highlighted how decomposability and transparency act at two different intersections: between syntax and semantics the former, between semantics and conceptual system the latter.

supports the interpretation that making something visible (spilling beans) is akin to making something known (revealing a secret). Similarly, CONTROL IS A BOUNDED REGION conceptualizes the act of controlling information as keeping it within a defined boundary, with the act of spilling signifying a loss of control and thus the escape of information beyond its intended confines. Therefore, transparency acts at the intersection of lexical semantics and conceptual system: the more the relationship between the frames is motivated (by metaphors, metonymies, encyclopedic knowledge Kovecses and Szabco, 1996; Carrol et al., 2018), the more transparent it is.

In spite of FrameNet's suitability for modeling idioms in a multi-layered way (thus fully respecting their complex nature), the English database currently includes only 35 idiomatic lexical units evoking 29 frames. Moreover, annotations based on real text are available for only 12 of these expressions. This means that considerable work is needed: first, to expand the database with new idioms; second, to add the suggested relations to investigate idioms' decomposability and transparency.

Nevertheless, it is worth recalling that one of Frame Semantics and FrameNet's core strengths lies in bridging meaning with experience and culture (Fillmore, 2006) via real text annotations. Idioms, as condensates of cultural, experiential knowledge (Colston, 2015), offer invaluable insights into this link. Enriching the database with more idioms would be an opportunity to leverage FrameNet's full potential in capturing the interplay between idioms, culture and cognition.

5. Conclusion

The present work focused on the distinction between idiom decomposability and transparency, two key variables for the analysis of the interplay between idioms' literal and figurative dimensions.

Leveraging a normed lexicon of Italian and English idioms, the impact of decomposability and transparency on idiom meaning recognition was analyzed. The results show that decomposability and transparency make different contributions across the two languages, and suggest that the two variables are distinct from each other. Further research is necessary to explore the motivations behind this cross-linguistic difference, for instance by focusing on the interrelationship among idiom knowledge, decomposability, and transparency.

Following the call for a "cognitive (psycho)linguistics" (Espinal and Mateu, 2010) in idiom investigation, the study leveraged empirical results to foster interdisciplinary research. In this spirit, it was illustrated how FrameNet and MetaNet are ideal computational tools for modeling idiom decomposability and transparency. Accordingly, this interdisciplinary approach is a first step to bridge the existing gap in idiom research between empirical psycholinguistic investigations, theoretical linguistic analyses, and practical computational applications.

6. Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK2636.

7. Ethical considerations and limitations

All participants of the cross-linguistic norming study for the creation of the idiom lexicon provided informed consent and were fairly compensated for their time and involvement.

As already pointed out in the text, one major limitation of the study is the high average degree of idiom objective knowledge in the dataset. Therefore, follow-up research based on controlling the degree of idiom knowledge through the creation of ad-hoc idiomatic expressions has been suggested.

One additional limitation stems from the very nature of norming studies, in which participants are asked to assess certain linguistic variables. Such a task requires some degree of training through very specific instructions. This can be cognitively demanding for participants, and data are elicited in a contrived manner. For this reason, more indirect methods for measuring decomposability and transparency should be devised for future research, possibly with idioms in context rather than in isolation.

The suggested idiom analysis should be extended to other and different types of idioms. Just as an example, there are idioms whose transparency is not justified by metaphors. In this case, the literal-figurative frame-to-frame relation would rather be motivated by metonymies or encyclopedic knowledge.

Finally, the inherently limited nature of resources such as FrameNet and MetaNet, which require much manual work, is a further limitation. It has already been mentioned that FrameNet would need additions of new idioms and relations between frames and between FEs to implement the analysis outlined. Yet, these limitations can be viewed as opportunities for pioneering systematic incorporation of more figurative language into the database.

8. Conflict of Interest Statement

The author has no conflicts of interest to declare.

9. Data and Code Availability

The normed lexicon of English and Italian idioms can be found in the repository of the University of Göttingen, and is accessible under the CC BY-NC 4.0 license. The R file with the code of the statistical analysis is available at this link.

10. Bibliographical References

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Nyssa Z Bulkes and Darren Tanner. 2017. “going to town”: Large-scale norming and statistical analysis of 870 american english idioms. *Behavior research methods*, 49:772–783.

Cristina Cacciari and Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of Memory and Language*, 27(6):668–683.

Cristina Cacciari and Patrizia Tabossi. 2014. *Idioms: Processing, structure, and interpretation*. Psychology Press.

Gareth Carrol, Jeannette Littlemore, and Margaret Gillon Dowens. 2018. Of false friends and familiar foes: Comparing native and non-native understanding of figurative phrases. *Lingua*, 204:21–44.

Herbert L. Colston. 2015. *Using Figurative Language*. Cambridge University Press.

D A Cruse. 1986. *Cambridge textbooks in linguistics: Lexical semantics*. Cambridge University Press, Cambridge, England.

J. Cooper Cutting and Kathryn Bock. 1997. That’s the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blendsß. *Memory & Cognition*, 25(1):57–71.

Dmitrij Dobrovol’skij. 2016. The notion of "inner form" and idiom semantics. *Études et travaux d’Eur’ORBEM*, 1(1):21–35.

M. Teresa Espinal and Jaume Mateu. 2010. On classes of idioms and their interpretation. *Journal of Pragmatics*, 42(5):1397–1411.

Charles Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di semantica*, 6:222–254.

Charles J. Fillmore. 2006. *Frame semantics*, page 373–400. Mouton de Gruyter.

Dirk Geeraerts. 1995. Specialization and reinterpretation in idioms. *Idioms: Structural and psychological perspectives*, 57:1–14.

Dirk Geeraerts. 2003. *The interaction of metaphor and metonymy in composite expressions*, page 435–468. Mouton de Gruyter.

Ferdy Hubers, Catia Cucchiarini, Helmer Strik, and Ton Dijkstra. 2019. Normative data of dutch idiomatic expressions: Subjective judgments you can bank on. *Frontiers in Psychology*, 10:1075.

Ray Jackendoff. 1995. The boundaries of the lexicon. *Idioms: Structural and psychological perspectives*, 133:165.

Z. Kovecses and P. Szabco. 1996. Idioms: A view from cognitive semantics. *Applied Linguistics*, 17(3):326–355.

George Lakoff and Mark Johnson. 2008. *Metaphors We Live By*. University of Chicago Press, Chicago, IL.

Andreas Langlotz. 2006. Idiomatic creativity. *Idiomatic Creativity*, pages 1–339.

Maya R Libben and Debra A Titone. 2008. The multidetermined nature of idiom processing. *Memory & cognition*, 36:1103–1121.

Daniel Lüdecke, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60):3139.

Diana Michl. 2019. Metonymies are more literal than metaphors: Evidence from ratings of german idioms. *Language and Cognition*, 11(1):98–124.

Rosa Elena Vega Moreno. 2005. Idioms, transparency and pragmatic inference. Technical report, UCL Working Papers in Linguistics, 17: 389–425, 2005. 25.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Irene Pagliai. 2023. Bridging the gap: Creation of a lexicon of 150 pairs of english and italian idioms including normed variables for the exploration of idiomatic ambiguity. *Journal of Open Humanities Data*.

Miriam R.L. Petruck. 2016. Introduction to metanet. *MetaNet*, 8(2):133–140.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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.

Manfred Sailer. 2021. Idioms. *Head-Driven Phrase Structure Grammar: The Handbook*, pages 777–809.

John R. Searle. 1975. *Indirect Speech Acts*, page 59–82. BRILL.

S Sprenger, W Levelt, and G Kempen. 2006. Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, 54(2):161–184.

Patrizia Tabossi, Lisa Arduino, and Rachele Fanari. 2011. Descriptive norms for 245 italian idiomatic expressions. *Behavior Research Methods*, 43:110–123.

Debra A. Titone and Cynthia M. Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics*, 31(12):1655–1674.

Wiltrud Wagner. 2021. *Idioms and Ambiguity in Context*. De Gruyter.

Bodo Winter. 2019. *Statistics for linguists: An introduction using R*. Routledge, London, England.

Stefanie Wulff. 2013. Words and Idioms. In Thomas Hoffmann and Graeme Trousdale, editors, *The Oxford Handbook of Construction Grammar*, pages 274–289. Oxford University Press.

11. Language Resource References

Pagliai, Irene. 2023. *Normed lexicon of English and Italian idioms*. GRO.data.

What GPT-4 Knows about Aspectual Coercion: Focused on ‘Begin the Book’

Seohyun Im, Chungmin Lee
Seoul National University
Seoul, Korea
{ish97, clee }@snu.ac.kr

Abstract

This paper explores whether Pre-trained Large Language Models (PLLMs) like GPT-4 can grasp profound linguistic insights into language phenomena such as Aspectual Coercion through interaction with Microsoft’s Copilot, which integrates GPT-4. Firstly, we examined Copilot’s understanding of the co-occurrence constraints of the aspectual verb “begin” and the complex-type noun “book” using the classic illustration of Aspectual Coercion, “begin the book.” Secondly, we verified Copilot’s awareness of both the default interpretation of “begin the book” with no specific context and the contextually preferred interpretation. Ultimately, Copilot provided appropriate responses regarding potential interpretations of “begin the book” based on its distributional properties and context-dependent preferred interpretations. However, it did not furnish sophisticated explanations concerning these interpretations from a linguistic theoretical perspective. On the other hand, by offering diverse interpretations grounded in distributional properties, language models like GPT-4 demonstrated their potential contribution to the refinement of linguistic theories. Furthermore, we suggested the feasibility of employing Language Models to construct language resources associated with language phenomena including Aspectual Coercion.

Keywords: Aspectual Coercion, GPT-4, Copilot

1. Introduction

This paper aims to explore what GPT-4, integrated into the Microsoft Copilot[1], knows about the Aspectual Coercion phenomenon and related linguistic theories.

According to Generative Lexicon theory (Pustejovsky, 1995), Type Coercion is a semantic operation that converts an argument to the type expected by a function, thereby preventing a type error. For instance, English verbs such as “begin” or “enjoy” typically take an event-type argument as their complement. In other words, the canonical semantic type of their complement is an event type, not an entity type. Although in the sentence “John began the book,” the complement “the book” is an entity type, American English native speakers generally accept this construction as both grammatically correct and semantically natural. Pustejovsky (1995) argues that this acceptance arises because speakers assume an activity related to the book. Formally, a main verb like “begin” or “enjoy” coerces the semantic type of its complement into an event type, aligning with the conventional interpretation of the sentence, thus rectifying any potential type error.

Aspectual Coercion specifically refers to Type Coercion by aspectual verbs such as “begin”, “continue”, “end”, or “finish” [2]. The language’s conventional interpretation coerces the argument’s meaning into an appropriate interpretation while forcibly aligning the argument’s type with an event type (Pustejovsky and Bouillon, 1995)[3].

One of the fundamental characteristics of language is the principle of linguistic economy, employing the minimal expression required to convey their intended message (Culicover and Jackendoff, 1990). This principle often entails omitting linguistic expressions denoting information that can be inferred implicitly, relying on shared conventions, common/world knowledge, or situational context within conversation. Instances such as Aspectual Coercion exemplify common language phenomena where implicit meanings are interpreted based on linguistic conventions, common sense, or contextual cues, whether linguistic or non-linguistic in nature.

When considering the principle of linguistic economy from a Natural Language Processing (NLP) perspective, it presents a quite difficult challenge. Traditional machine learning-based NLP methodologies often entail enumerating various types of inferences, exemplified by research in Natural Language Inference (NLI), and constructing training datasets for these inferences (MacCartney et al., 2006; Im and Pustejovsky, 2009). However, with phenomena like Aspectual Coercion, the implicit meaning becomes inherently ambiguous, and the specific interpretation of a type-coerced sentence can fluctuate depending on diverse contextual factors. Furthermore, disparate interpretations may hold precedence in different contexts, without necessarily excluding alternative understandings. As a result, addressing challenges posed by Aspectual Coercion remains among the most arduous tasks within traditional machine learning-based NLP.

The emergence of powerful Natural Language Generation (NLG) models like OpenAI’s GPT-4 (Josh et al., 2023) signifies a remarkable transformation in

[1] Refer to [Microsoft Copilot for Microsoft 365 overview | Microsoft Learn] to learn about Copilot.

[2] Michaelis (2004, 2022) considers aspectual type shift coerced by certain expressions in a sentence as aspectual coercion (e.g., “I was outside twice”). In this paper, we adopt Aspectual Coercion as proposed by Pustejovsky and Bouillon (1995).

[3] Im and Lee (2013) discuss the phenomenon of type coercion associated with the light verb “ha-” (‘do’) in Korean, drawing upon the explanation of aspectual coercion observed with the English verb “begin” within the framework of Generative Lexicon theory.

The Workshop on Cognitive Aspects of the Lexicon, pages 56–67
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

the landscape of NLP. This shift marks a notable turning point, presenting unprecedented opportunities.

Nevertheless, we question whether Pre-trained Large Language Models (PLLMs) like GPT-4 can offer a linguistic theoretical explanation of the Aspectual Coercion Phenomenon. This skepticism arises from the models' foundation on the Distributional Hypothesis (Harris, 1954), which posits that word meanings are described by their contextual usage. Distributional semantics solely reveals the distributional characteristics of words without delving into the cognitive or conceptual aspects of why these properties emerge within the lexicon of human language.

In this context, our aim is to investigate whether a PLLM such as GPT-4 can accurately and completely provide linguistic explanations for Aspectual Coercion. Additionally, we examine the intricacies of designing prompts that effectively enable the extraction of precise linguistic insights about Aspectual Coercion.

In sections 2.1 and 2.2, we discuss the semantic nature and co-occurrence constraints of the aspectual verb "begin" and its direct object "book". This analysis is grounded in the framework of the Generative Lexicon Theory and Type Composition Logic (Asher and Pustejovsky, 2013). In section 2.3, we investigate GPT-4's understanding of the semantic attributes of the verb "begin" and the noun "book" through prompting. More specifically, we employ Microsoft's Copilot, an AI-powered assistant which leverages GPT-4. Section 3 consolidates the previous discussions and scrutinizes the Aspectual Coercion phenomenon that emerges during the semantic composition of "John," "began," and "the book." We particularly explore Copilot's interpretations of Aspectual Coercion sentences in section 3.2. Lastly, Section 4 summarizes our research findings and presents conclusions.

2. The Aspectual Verb "Begin" and the Noun "Book"

2.1 The Aspectual Transitive Verb "Begin"

Aspectual verbs such as "begin", "continue", and "finish" play a crucial role in encoding distinctions related to the temporal aspect of actions or events. In particular, the English aspectual verb "begin" is frequently used to indicate the start or commencement of an action, process, or event. This verb, when used transitively, accepts various types of complements: A direct object noun (e.g., "She began her <u>study</u> an hour ago."), a "to"-infinitive (e.g., "He began <u>to read</u> the novel a week ago."), and a gerund (e.g., "He began <u>writing</u> an essay a month ago.").

In the examples provided, the direct object "her study" is an event type argument, as do "to read the novel" and "writing". Conversely, the sentence "She began the rock" is semantically awkward because the direct object "the rock" represents an entity type rather than an event. American English native speakers typically do not perceive such a sentence as semantically natural.

The first constraint in complement selection of the aspectual verb "begin" is that:

(1) **Complement Selection Constraint 1**: It typically requires an <u>event or action-type</u> argument as its canonical complement. It cannot accommodate other types of arguments. (Pustejovsky, 1995)

However, the verb cannot take all kinds of event-type argument. Given that "begin" refers to the initial phase of an event or action that involves a duration of time, the argument of the verb cannot depict an event that ends in a very short moment. For example, the sentence "She began to <u>blink</u> her eyes one time" is not grammatically valid. Similarly, the sentence "She began to <u>arrive</u> at school" is also considered ungrammatical. This leads to the second constraint of the verb's complement selection:

(2) **Complement Selection Constraint 2**: It can NOT take an <u>achievement or punctual</u> event-type complement. (Im and Lee, 2015)

To summarize, the aspectual verb "begin" should take an event-type argument as its complement. However, the event-type complement should not be an achievement or punctual event type.

2.2 The Dot Object Noun "Book"

In this section, we propose a slight adjustment to the semantic type of the English noun "book" as proposed in Im and Lee (2015). Specifically:

(3) The semantic type of "book"
text_info·bound_paper·goods

This semantic type represents "book" as a Dot Object comprising the subtypes text_info, bound_paper, and goods[4]. Dot objects, introduced by Pustejovsky (1994), refers to objects characterized by a complex type, exemplified by the semantic type of "book" outlined in (3) above.

Such a complex type of a noun facilitates predications that are licensed over one of the three dot element types (Asher and Pustejovsky, 2013). Consequently, each subtype in (3) implies a set of predicates that can co-occur with the noun "book" as detailed below.

- **text_info** → {read, write, translate, publish, digitize, evaluate, underestimate, criticize, create, …}
- **bound_paper** → {take, carry, put, place, position, pack, tear, burn, weight, borrow, own, have, lend, …}

[4] We consider that the semantic type of "book" presupposes a paper book in this paper. However, we notice that a book can represent special types of books such as an audiobook or a picture book. In future work, we need to update the semantic type of "book" based on corpus analysis. Language Models like GPT-4 can be helpful for this task.

- **goods** → {buy, sell, promote, market, advertise, …}

Even more intriguingly, the semantic type of “book” implies a set of predicates that cannot co-occur with the noun. For instance, the noun “book” cannot be a direct object of the verbs presented in (4).

(4) Verbs that cannot co-occur with “book”
{eat, drink, wear, sing, mow, bake, …}

To sum up, the complex type of “book” dictates which predicates can co-occur with the word and which cannot.

2.3 Copilot’s Understanding of the words “Begin” and “Book”

In this section, we pose inquiries to Copilot regarding the lexical semantics and co-occurrence constraints of the aspectual verb “begin” and the complex-type noun “book[5][6]. The first question in Prompt 1 pertained to the direct object selection constraints of the aspectual verb “begin”.

> **Question**: What are the lexical semantic constraints for a complement of the transitive verb ‘begin’?
> **Answer**: … Aspectual verbs like ‘begin’ have been assumed to <u>select exclusively for eventive complements</u>, such as in the sentence ‘John began reading the book’. …

Prompt 1 Complement selection constraints of the aspectual verb 'begin'

Prompt 1 demonstrates that Copilot recognizes that aspectual verbs like “begin” typically require event-type complements. However, it does not demonstrate knowledge of the more specific constraint that the verb “begin” does not accept achievement or punctual event-type arguments as its complement.

To compile a list of the most common verbs that can be inserted into the structure “John began to _____ the book,” we employed Prompt 2 as follows:

> **Question**: Please list the most common verbs that can be inserted into “John began to _____ the book.” in order of frequency.
> **Answer**: … 1. Read, 2. Write, 3. Study, 4. Open, 5. Close, 6. Finish, 7. Review, 8. Mark, 9. Cover, 10. Illustrate, 11. Edit, 12. Publish, 13. Sell, 14. Buy, 15. Bind, 16. Sign, 17. Dedicate, 18. Translate, 19. Print, 20. Annotate.

Prompt 2 The most common twenty transitive verbs that co-occur with 'begin' and 'the book'

We tasked Copilot with listing verbs in order of frequency. From Copilot’s output, we found that the verb “read” is the most common verb between the aspectual verb “begin” and the noun “book”. The second most common candidate is “write.” This highlights the advantage of Language Models in situations where we cannot determine, based solely on linguistic knowledge, which of the verbs is the most common verb to co-occur with “book”.

In addition, Copilot noted that these twenty verbs represent the most common activities or events of duration that individuals typically undertake with a book. This suggests that Copilot emphasized the co-occurrence of the verbs with the noun “book” rather than with the aspectual verb “begin”. In essence, Copilot does not consider the constraints of semantic type composition when words are combined to form a sentence.

The sixth verb, “finish,” in the list is not directly related to any subtype of the noun “book.” It also requires a form of aspectual coercion to co-occur with the noun “book.” However, GPT-4, relying on word distribution, simply lists the most frequent verbs that take “book” as their direct object, regardless of the type composition constraints of the noun “book” and its governing verbs.

Next, we employed another prompt to inquire about verbs that are not suitable for insertion into the sentence. In Prompt 3, Copilot presented the twenty verbs as the least likely options to co-occur with “begin” and “the book”. Furthermore, it provided an explanation for selecting this list of verbs: These verbs denote actions that are not commonly linked with a book or reading.

> **Question**: Please list the 20 transitive verbs that cannot be inserted into ‘John began to ____ the book.’ with the reason.
> **Answer**: 1. Swim, 2. Jump, 3. Run, 4. Fly, 5. Breathe, 6. Eat, 7. Drink, 8. Sleep, 9. Dance, 10. Sing, 11. Laugh, 12. Cry, 13. Sneeze, 14. Walk, 15. Drive, 16. Climb, 17. Smile, 18. Frown, 19. Blink, 20. Sweat
> These verbs are awkward in this context because they denote <u>actions that are not typically associated with a book or reading</u>. …

Prompt 3 The most improbable transitive verbs that co-occur with "begin and "the book"

Copilot solely concentrated on actions that are not typically associated with a book, disregarding verbs that cannot co-occur with the aspectual verb “begin” even though they can take “the book” as their complement. For example, the verb “obtain” is an achievement verb and cannot be combined with “begin”, despite its compatibility with “the book” as a complement. However, Copilot incorrectly indicated that the sentence “John began to obtain the book.” is semantically acceptable as shown in Prompt 4.

[5] Microsoft Bing AI Copilot has three modes of chatting: more balanced, more precise, and more creative. In this section, we utilized the more balanced mode for our interactions.

[6] For now, we include only essential content in prompting example boxes due to space limitations. All prompts containing full content will be provided in the Appendix when we have the opportunity to publish a camera-ready version of our paper.

Sentence: John began to obtain the book.
Question: Is the sentence semantically acceptable?
Answer: Yes, the sentence “John began to obtain the book.” is semantically acceptable. The verb “obtain” means to come into possession of something, so this sentence suggests that John started the process of getting the book. …

Prompt 4 Question regarding the possibility of co-occurrence of 'obtain' and 'begin'

From the prompts used, it seems that Copilot may not possess the understanding that the aspectual verb “begin” cannot be combined with achievement or punctual event-denoting verbs. To verify Copilot’s comprehension of the properties of achievement verbs, we posed a question using an alternative prompt.

Sentence: John obtained the book for three hours.
Question: Is the sentence semantically acceptable?
Answer: No, the sentence “John obtained the book for 3 hours” is not semantically acceptable. The verb “obtain” implies a change of possession, which is typically an instantaneous action rather than one that extends over a period of time. Therefore, it’s unusual to use “for three hours” with “obtain”. ...

Prompt 5 Is Copilot knowledgeable about the lexical semantic properties of achievement verbs?

Copilot’s response indicates its awareness that achievement verbs like “obtain” cannot co-occur with a prepositional phrase denoting time duration (e.g., “for three hours”). Therefore, we conclude that Copilot lacks the knowledge that the aspectual verb “begin” cannot accept achievement or punctual event-denoting verbs as its complement.
Furthermore, Copilot’s listing of the verbs that can co-occur with “book” does not align with the semantic type of the noun. We used another prompt to confirm this discrepancy.

Question: What is the semantic type of the English noun “book”?

Answer: The semantic type of the English noun “book” is associated with the category of “things” or “objects” …

Prompt 6 Semantic Type of the English Noun "book"

Based on Copilot’s answer in Prompt 6, we argue that it cannot provide deeper linguistic knowledge related to the complex type of “book” described in section 2.2. We asked Copilot more specifically using the question “What is the semantic type of “book” in Generative Lexicon Theory or Type Composition Logic?”. However, Copilot could not identify “book” as a complex type (dot object) noun. In conclusion, GPT-4 lacks deeper knowledge about linguistic constraints regarding composition of words.

3. Composition and Interpretation of the Sentence “John began the book”

3.1 Interpretation of the Sentence “John began the book”

Aspectual Coercion refers to the process in which an aspectual verb changes the semantic type of its direct object when the object does not meet the constraints imposed by the aspectual verb. Consequently, the complement undergoes Type Shifting to a suitable type. The aspectual verb “begin”, which requires an event type expression as its direct object, coerces the semantic type of “book” – originally a complex type text_info·bound_paper·goods – into an event type.

Im and Lee (2015) proposed constraints on the aspectual coercion of the verb “begin” in the sentence “John began the book,” drawing from the work of Pustejovsky and Bouillon (1995) and the constraints discussed in section 2:

A. Only in the control construction of “begin”
B. When the subject is animate
C. Only when the missing predicate belongs to a process or an accomplishment type verb class (neither a punctual nor an achievement)
D. And only when the missing predicate is a two-place verb which takes a subject and an object.

The examples provided in (5-8) show violations of the conditions of Aspectual Coercion.

(5) The book begins with the word. → *The book begins the word. (constraint A)
(6) *The rock began reading the book. → *The rock began the book. (constraint B)
(7) *John began losing the book. → *John began the book. (constraint C)
(8) John began giving me books. → *John began me books. (constraint D)

Inversely, the constraints of Aspectual Coercion suggest the limitations on candidates for the implicit predicate that is not explicitly encoded on the surface. This predicate should be inferred by native speakers of the language based on context, linguistic conventions, and common sense or world knowledge when interpreting the Aspectual Coercion construction.

Firstly, when the words “John,” “began,” and “book” come together to form a sentence, their semantic types also interact. In essence, the combination of “began” and “the book” necessitates an intersection of the verbs permitted by “began” and those governing “the book.” Subsequently, this intersection set intersects with the set of verbs permitted by “John.” For example, “began the book” can be interpreted as “began reading the book,” but it cannot be understood as “began losing the book” or “began swimming the book.” The subject “John”, which gives no specific information, imposes no limitation on the possible interpretations of the sentence. Therefore, the sentence can be interpreted as follows:

“John began to **read/write/publish/study/…** the book.”

The most common interpretation in this case depends on linguistic conventions and common knowledge in English. Prompt 2 showed that reading interpretation is most common.

Secondly, with a more specific linguistic context provided in the sentence, the preferences for interpretations change.

Sentence: The writer began the book.
Preferred interpretation: The writer began to write the book.

The preferred interpretation leans towards writing, given that the subject "the writer" provides more specific context indicating that the individual initiating the book is indeed a writer. It is important to note that all possible interpretations derived from "began the book," including reading interpretation, are still considered, despite the changes in preferred ratings.

Thirdly, when a specific situational context is given, the preferred interpretation of the sentence also changes.

Sentence: John began the book.
Situational context: John went to the shop that sells chocolates. He bought a chocolate box with various shapes of chocolate including a book-shaped chocolate.
Preferred Interpretation: John began eating the book.

In this case, the preferred interpretation of the sentence is 'John began eating the book' as mentioned above. More accurately, it is interpreted as John began eating the book-shaped chocolate[7]. We mention that an interpretation like 'John began reading the book' is possible, but "the book" does not denote the book-shaped chocolate but a book the speaker or John considers in their mind.

In summary, we presented comprehensive constraints of Aspectual Coercion and demonstrated that the preferred interpretations of Aspectual Coercion sentences depend on contextual cues. We emphasize that although preferred interpretations change according to contextual information, all possible interpretations remain.

3.2 Copilot's Interpretation of the Aspectual Coercion Sentence

To verify if Copilot possesses knowledge about aspectual coercion of the verb "begin" and the interpretation of the construction, we posed various questions through prompting. Initially, we simply asked the meaning of the sentence "John began the book."

Copilot offers three sessions: the more balanced session for casual, informed chat; the more precise session for concise chat aimed at helping you discover facts; and the more creative session for original and imaginative chat. We compared interpretations of the Aspectual Coercion sentence across the three modes of Copilot. As observed in Prompt 7, all three modes showed no difference in their answers, providing interpretations such as reading or writing the book. The more creative session added another interpretation, such as "John opened the book," which could be considered an overgeneration.

> **Sentence**: John began the book.
> **Question**: What is the meaning of the sentence?
>
> **More balanced mode answer**: The sentence "John began the book" typically means that John started reading a book or started writing a book. The exact meaning depends on the context….
> **More precise mode answer**: … when no specific context is provided, the most common understanding is that John started reading or writing the book. …
> **More creative mode answer**: The sentence 'John began the book' conveys the idea that John initiated the act of starting to read or write the book.

Prompt 7 Interpretation in the More Balanced session

When we inquired about the most common interpretation of the sentence with no specific context, Copilot gave the exact answer we intended, "John started reading a book," only in the more precise session. However, it also mentioned the possibility of a writing interpretation and emphasized the importance of context (Prompt 8).

> **Sentence**: John began the book. **Question**: What is the most common interpretation of the sentence when no specific context is given in English?
>
> **Answer**: The most common interpretation of the sentence "John began the book" when no specific context is given in English is that John started reading a book. …

Prompt 8 The most common interpretation of the Aspectual Coercion sentence when no specific context is provided

As discussed in section 3.1, the Aspectual Coercion sentence remains ambiguous even though the most preferred interpretation is "John began to read the book." Therefore, we requested Copilot to provide a list of twenty possible interpretations of the sentence when no specific context is given. For this question, we selected the more precise session. We only display answers from the more precise session from now on.

> **Sentence**: John began the book.
> **Question**: Please list the 20 possible interpretations of the sentence.
>
> **Answer**: Sure, here are 20 possible interpretations of the sentence "John began the book": 1. John

[7] Even without this specific context, the change of the subject to "the goat (non-human animate agent)" triggers an eating interpretation as the preferred interpretation of the sentence "The goat began the book."

started reading a book. 2. John started writing a book. 3. John opened the book to start reading. 4. John started studying the book for academic purposes. 5. John began narrating the book to someone else. 6. John started translating the book into another language. 7. John started illustrating a book. 8. John began binding the book. 9. John started selling the book. 10. John began a book club focused on the book. 11. John started a journey that the book is about. 12. John began to critique the book. 13. John started to adapt the book into a movie or play. 14. John began to memorize the content of the book. 15. John started to audio-record the book. 16. John began to teach the content of the book. 17. John started to convert the book into braille. 18. John began to digitize the book. 19. John started to restore the old book. 20. John began to research about the book's background. …

Prompt 9 Twenty possible interpretations of the sentence

Prompt 9 suggested verbs denoting common activities that people engage in with a book, including reading and writing. Interestingly, it provided various types of interpretations beyond just recovering the implicit verb that governs "the book" as its direct object. For instance, Prompt 9 suggested "began a book club focused on the book." This finding raises questions about previous linguistic research on Aspectual Coercion phenomena and the interpretations of the construction. Perhaps we should develop more elaborate linguistic explanations that encompass various types of interpretations, rather than solely focusing on recovering implicit verbs.

To examine whether Copilot is aware of the linguistic theoretical constraints for interpreting the Aspectual Coercion sentence, we inquired about the criteria it used to select the interpretations it showed.

Question: What were the criteria for selecting the 20 interpretations?
Answer: Here are the criteria I used:

1. Different Actions: …
2. Different Roles: …
3. Different Contexts: …
4. Feasibility: …
5. Variety: …

Remember, the exact interpretation of a sentence often depends on the context in which it's used. Without additional context, multiple interpretations could be valid.

Prompt 10 Copilot's criteria for selecting twenty possible interpretations

In Prompt 10, regarding different actions, Copilot suggests that the verb "began" can imply various actions such as reading, writing, translating, selling, etc. Each action provides a different interpretation of the sentence. Copilot also considered different roles the subject "John" can take in relation to the book – reader, writer, seller, teacher, etc. Third, the Copilot considered that the context in which the sentence is used can drastically change its meaning. For example, in a publishing context, "John began the book" could mean he started writing it. In a bookstore, it could mean he started selling it. Copilot mentioned feasibility. The interpretations should be plausible scenarios that could realistically occur in everyday life. Finally, Copilot considered as many different actions and roles as possible to provide a wide range of interpretations. Again, the criteria suggested by Copilot do not encompass the constraints provided by the verb "begin", instead focusing solely on the meaning of "book".

When a specific context is provided, the most appropriate interpretation of the sentence depends on the context. We wrote a specific context together with the aspectual coercion sentence and a question about its meaning in Prompt 11.

Context: John is an editor who works for a publishing company. **Sentence**: John began the book. **Question**: What is the most appropriate interpretation of the sentence when the context is given ?
Answer: Given the context that John is an editor working for a publishing company, the most appropriate interpretation of the sentence "John began the book" would be that John started editing the book. This interpretation assumes that "the book" refers to a manuscript or a draft that John is working on as part of his job as an editor. …

Prompt 11 Interpretation of the sentence when a specific context is provided

Because the context gave a specific information about "John," Copilot suggested the most appropriate interpretation of the sentence as "John started editing the book". However, the interpretation cannot exclude the reading or writing or even other possible interpretations of the sentence. We check if Copilot knows about the fact by using another prompt.

Question: Under the context, is the only interpretation you gave possible? Please give me other possible 10 interpretations of the sentence when the context is given.
Answer: Given the context that John is an editor working for a publishing company, here are 10 other possible interpretations of the sentence "John began the book":
1. John started to review the book for potential publication. 2. John began to format the book for printing. 3. John started to plan the layout of the book. 4. John began to proofread the book. 5. John started to assign the book to a specific genre or category. 6. John began to write a summary or blurb for the book. 7. John started to coordinate with the author about revisions for the book. 8. John began to plan the marketing strategy for the book. 9. John started to discuss the book with other editors or staff. 10. John began to prepare the book for digital publication.

Prompt 12 Ten other possible interpretations

In the given context, when asked for ten additional interpretations other than the translation provided by the Copilot, it suggested the interpretations in Prompt 12. Most of these interpretations were related to tasks associated with editors working for publishing companies. The Prompt provided interpretations with various types of structure, rather than simply recovering elided predicates. This fact hints to us that we need to generalize and explain the kinds of interpretations in terms of linguistic theory.

We were curious if Copilot knew, even in the context of John being an editor, that the sentence can still be "John began reading the book." Therefore, we asked about it with another prompt.

> **Question**: Is it impossible to interpret the sentence as 'John began reading the book' under the context?
> **Answer**: No, it's not impossible. … it's still plausible that "John began the book" could mean "John began reading the book". … Context is key, and without more specific information, multiple interpretations can still be valid. … Without additional context, multiple interpretations could be valid.

Prompt 13 Is reading interpretation possible under the context in Prompt 12?

The response from Copilot to the question in the above prompt reflects that it recognizes that while the interpretation contextualized in the given context is the most likely interpretation, there is openness to the possibility of other interpretations.

Another question we asked is about impossible interpretations of the Aspectual Coercion sentence when no specific context is given. We show only five interpretations in Prompt 14. The Prompt explains the reason that these examples are impossible interpretations as follows: These interpretations are impossible because they involve using a book in ways that defy the laws of physics or the nature of a book.

> **Sentence**: John began the book. Question: Please list the 20 impossible interpretations of the sentence.
>
> **Answer**: 1. John started to eat the book. 2. John began to use the book as a spaceship. 3. John started to wear the book as a hat. 4. John began to play the book like a guitar. 5. John started to drink the book. …

Prompt 14 Impossible interpretations

Now, we asked a question about the same aspectual coercion sentence with a specific context in the following prompt.

> **Context**: John's father gave a chocolate box in which there are various shapes of chocolates. One of them is a book-shape. **Sentence**: John began the book. **Question**: What is a common interpretation of the sentence when the context is given?
> **Answer**: Given the context, a common interpretation of the sentence could be that John started eating the chocolate that is shaped like a book. This interpretation assumes that "the book" refers to the book-shaped chocolate in the context provided. Remember, the exact interpretation of a sentence often depends on the context in which it's used. Without additional context, multiple interpretations could be valid.

Prompt 15 Common interpretation of the sentence under a specific situational context

In the more precise Copilot mode, it gave the exact answer that we wanted when we gave a specific context. A common interpretation of the sentence is that John started eating the chocolate that is shaped like a book, assuming "the book" refers to the book-shaped chocolate. It is very interesting compared to Copilot's answer to the question regarding the impossible interpretations of the sentence when no specific context is given. In the case, Copilot suggested John started to eat the book as one of the impossible interpretations. On the contrary, Copilot showed that the sentence is interpreted as "John began eating the book" when the appropriate specific context is provided[8].

Finally, we propose the possibility of building a lexical resource that systematically organizes information about Aspectual Coercion and the interpretation of Aspectual Coercion sentences through prompting. In the following prompt, we listed aspectual verbs and then described the task Copilot should perform as kind of pseudo code. Copilot performed well as we intended. It created Aspectual Coercion sentences with each of the verbs in the list, "Words." And then it provided five possible interpretations for each aspectual coercion sentence.

> **Words** = ['begin', 'finish'] **Sentence**: John _____ the book. **Question**: For each word in words, please make a sentence by inserting each word into ____. And then please list 5 possible interpretations of each sentence in order of frequency.
> **Answer**: Here are the possible interpretations for each word:
> **1. John began the book.**
> - John started writing a novel, perhaps sitting at his desk with a blank page before him.
> - John initiated the process of creating a literary work, whether fiction or non-fiction.
> - John embarked on the journey of authorship, with ideas swirling in his mind.
> - John opened a new chapter in his creative endeavors.
>
> **2. John finished the book.**

[8] In an imagined or fictional contextual scenario, a more intriguing issue of interpretation arises. For instance, when we present the sentence "The lady in the picture began a song" and inquired about its meaning to Copilot, it did not provide the intended answer but instead suggested longer sentences with more specific contextual details, indicating that the context is fictional or imagined. Copilot's response to the question implies that it cannot interpret the sentence as we, human being, can.

- John completed reading a captivating novel, turning the final page with satisfaction.
- John concluded writing a manuscript, typing the last sentence and feeling a sense of accomplishment.
- John reached the end of a literary work, whether it was a short story or an epic saga.
- John wrapped up the book project, editing and revising until it was polished.
- John closed the cover of a well-loved book, reflecting on the journey it took him on.

Remember, context matters, and the intended meaning depends on the surrounding information!

Prompt 16 Building a lexical resource about Aspectual Coercion

4. Related Work

The coercion phenomenon has been widely studied in theoretical, psychological, and computation linguistics. From the perspective of theoretical linguistics, coercion challenges to traditional semantic compositionality theory (Asher, 2015; Pustejovsky & Batiukova, 2019, Rambelli et al., 2020). Pustejovsky (1995) suggested that the superficial violation of compositionality can be resolved by recovering an implicit event-denoting verb based on the qualia of the noun in the complement NP (e.g., qualia of "book" in "begin the book"). Therefore, the source of interpretation of the coercion construction lies within our lexicon.

Alternatively, Zarcone et al. (2011, 2012, 2013, 2017) proposed, based on several experimental and computational linguistic studies, that the source of interpretation is generalized event knowledge, that is, pragmatic and world knowledge. In particular, Zarcone et al. (2017) explored the interaction between the semantic type of the object (event vs. entity) and the typicality of the covert event ("the author began a book" → "writing") during the processing of coercion construction by employing a self-paced reading study. This interaction demonstrates the combined influences of verb-driven type preferences and generalized real-world event knowledge during language comprehension.

Psycholinguists and cognitive scientists are interested in determining whether there are indicators that could demonstrate the presence of coercion and implicit eventive expression by revealing extra processing costs during online sentence comprehension in cognitive research on human sentence processing (McElree et al., 2001; Traxler et al., 2002). Another intriguing question concerns the origins of the extra processing costs. One possible explanation is the retrieval of an event sense of the complement (e.g., "began **reading** the book"), while another is the relative unpredictability of the complement noun. Delogu et al. (2017) argue that the cost largely corresponds to the surprisal associated with the complement noun. Gu (2022) utilized surprisal estimates at critical sentence positions to investigate how Large Language Models (LLMs) respond to implicit meaning such as type coercion. They demonstrated that surprisal estimates in Language Models (LMs) reflect the difficulty involved in recovering the covert meaning.

Computational studies on type coercion focus on interpreting coerced sentences, aiming to identify potential covert event candidates for the complement argument in aspectual coercion sentences. Representative computational models for interpreting type coercion sentences include probabilistic, distributional, and Large Language Model (LLM)-based models. First, the probabilistic model, as propsed by Lapata and Lascarides (2003), considers the interpretation of a coercion sentence as a joint distribution P(subject (s), coercion verb (v), the object (o), covert event (e)). The preferred interpretation of the coercion sentence is then the event that maximizes P(s, v, o, e). Second, Zarcone et al. (2012, 2013) introduced a distributional semantic model that identifies the covert event as the one with the highest thematic fit with the complement in the coercion sentence. Additionally, Chersoni et al. (2019) proposed the Structured Distributional Model (SDM) that integrates word embeddings with formal semantics. This model incorporates the psycholinguistic research findings that generalized knowledge about events stored in semantic memory plays a crucial role in sentence comprehension.

Recent studies have employed LLMs to address the challenge of interpreting coercion constructions. Ye et al. (2022) introduced a BERT-based dense paraphrasing model, which combines paraphrasing and decontextualization (Choi et al., 2021). Their results surpassed those of previous statistical and distributional models, suggesting that while coercion construction interpretation remains challenging even for LLMs, model performance can be enhanced by fine-tuning LLMs through dense paraphrasing. In a study by Rambelli et al. (2020), various models, including probabilistic, distributional, and LLM-based ones, were compared in terms of their effectiveness in interpreting coercion sentences. The findings indicate that the top-performing LLM-based models and some traditional distributional models exhibit comparable performance. Despite the diverse computational approaches to modeling the interpretation of type coercion constructions, experimental results suggest that type coercion remains a challenging phenomenon for computational modeling.

5. Conclusion

In this paper, we aimed to investigate the understanding of phenomena such as Aspectual Coercion by Pre-trained Large Language Models (PLLMs) like GPT-4 and their ability to accurately interpret Aspectual Coercion sentences. To achieve this, we prompted Microsoft's Copilot, which incorporates GPT-4, with various questions.

In section 2.1, we presented the lexical meaning and compositional constraints of the aspectual verb "begin" and the complex noun "book." Our interaction with Copilot in section 2.2 revealed its limitations in providing deeper linguistic insights about the lexical semantic properties of the aspectual verb “begin” and the noun “book.” Specifically, while it could generate a list of verbs that can co-occur with “begin” and “book” based on its ability of extracting distributional properties of words, it struggled to explain the underlying rationale behind its suggestions with deeper linguistic knowledge.

Section 3 explores the interpretation of Aspectual Coercion sentences containing “began the book.” When interpreting the Aspectual Coercion construction, the retrieval of missing information is deduced from the intersection set of candidates that each word allows for in the sentence when forming the Aspectual Coercion sentence. Copilot generally provided accurate interpretations of Aspectual Coercion sentences and demonstrated an understanding that such sentences are ambiguous and can be interpreted in various ways. Additionally, it offered interpretations beyond merely recovering omitted verbs, presenting a range of interpretations. This suggests that linguistic theoretical research on Aspectual Coercion should offer more sophisticated explanations that encompass such diverse interpretations.

Next, we posed several questions to determine whether Copilot recognizes that the preferred interpretation of the Aspectual Coercion construction changes when specific contexts are provided and whether it remains ambiguous. As a result, Copilot demonstrated its awareness that the preferred interpretation changes depending on the context and offered interpretations that are suitable for the context. Additionally, we confirmed that Copilot understands that while the preferred interpretation may change, the sentence remains ambiguous. Furthermore, Copilot indicated impossible interpretations when no context was provided but also demonstrated an understanding that these interpretations can change to possible ones when specific contexts are given. Finally, we discussed the potential for creating a lexical resource that offers insights into Aspectual Coercion through prompts using pseudo-code.

Through our research, we have affirmed that Language Models such as GPT-4 offer advantages in extracting preferred interpretations of aspectual coercion sentences. This is because the LLMs are fundamentally data-driven and capture the distributional patterns of a language, enabling them to effectively discern and generate contextually appropriate interpretations. However, they face difficulties in capturing more profound linguistic or conceptual knowledge about aspectual coercion.

Copilot provides slightly different answers each time it is asked the same question. Therefore, it is somewhat risky to determine Copilot’s understanding of deeper linguistic knowledge based on a single inquiry. Additionally, there are various modes, and there is also an option in the personal settings to decide whether Copilot will remember the queries and responses exchanged. Therefore, to accurately assess Copilot’s linguistic knowledge, it is necessary to make diverse and repeated attempts. These issues will be addressed in future research, and efforts will be made to use the GPT-4 API provided by OpenAI to attempt more comprehensive prompt-based learning. Furthermore, comparing it with Chat-GPT would be interesting.

We are planning to expand our research to include other LLMs. Additionally, we are considering the integration of linguistic knowledge into LLMs as a means to enhance the explainability of LLMs concerning Aspectual Coercion. It is hoped that this study will contribute to research on the interaction between Pre-trained Large Language models and linguistic theories. This interdisciplinary approach has the potential to enrich both fields and pave the way for further advancements in natural language understanding and modeling.

6. Bibliographical References

Asher, N. (2015). Types, Meanings and Coercion in Lexical Semantics. *Lingua*, 157: 66-82.

Asher, N. and Pustejovsky, J. (2013). A Type Composition Logic for Generative Lexicon. J. Pustejovsky et al. (eds.) *Advances in Generative Lexicon Theory*, Text, Speech and Language Technology 46, DOI 10.1007/978-94-007-5189-7_3, Springer Science+Business Media Dordrecht 2013.

Culicover, P. and Jackendoff, R. (1990). Economy and the grammar of the clause. *Language*, 66(4): 761-815.

Gu, Y. (2022) Measure More, Question More: Experimental Studies on Transformer-based Language Models and Complement Coercion. *arXiv preprint arXiv:2212.10536 (2022)*.

Harris, J. (1954). Distributional structure. *Word* 10(2-3): 146-162.

Im, S. and Lee, C. (2013). Combination of the verb HA- ‘do’ and entity type nouns in Korean: A Generative Lexicon Approach. J. Pustejovsky et al. (eds.) *Advances in Generative Lexicon Theory*, Text, Speech and Language Technology 46, DOI 10.1007/978-94-007-5189-7_9, Springer Science+Business Media Dordrecht 2013.

Im, S. and Lee, C. (2015). “Begin the book”: A developed analysis of Type Coercion based on Type Theory and conventionality. The ESSLLI *Proceedings of the TYTLES workshop on Type Theory and Lexical Semantics*.

Im, S. and Pustejovsky, J. (2009). Annotating event implicatures for textual inference tasks. Rumshisky, A. and Calzolari, N. (eds.) *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*.

Josh, A. et al. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

Lapata, M. and Lascarides, A. (2003). A Probabilistic Account of Logical Metonymy. *Computational Linguistics*, 29(2): 261-315.

MacCartney, B., Grenager, T., de Marneffe, M.-C., Cer, M. and Manning, C. D. (2006). Learning to recognize features of valid textual entailments. *North American Association for Computational Linguistics (NAACL)* 2006.

McElree, B., Traxler, M. J., Pickering, M. J., Seely, R. E., and Jackendoff, R. (2001). Reading Time

Evidence for Enriched Composition. *Cognition*, 78: B17-B25.
Michaelis, L. (2004). Type Shifting in Construction Grammar: An Integrated Approach to Aspectual Coercion. *Cognitive Linguistics* 15-1 (2004): 1-67.
Michaelis, L. (2022). Aspectual Coercion and Lexical Semantics Part 1: Using Selection to Describe the Interaction between Construction and Verb Meaning. *Cognitive Semantics* 8 (2022): 383-408.
Pustejovsky, J. (1994). Semantic typing and degrees of polymorphism. C. Martin-Vide (ed.) *Current Issues in mathematical linguistics*. Holland: Elsevier.
Pustejovsky, J. (1995). *The Generative Lexicon*. The MIT Press.
Pustejovsky, J. and Batiukova, O. (2019). *The Lexicon*. Cambridge University Press.
Pustejovsky, J. and Bouillon, P. (1995). Aspectual coercion and logical polysemy. *Journal of semantics* 12(2): 133-162.
Rambelli, G., Chersoni, E., Lenci, A., Blache, P., and Huang, C.-R. (2020). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 224-234 December 4-7, 2020.
Traxler, M. J., Pickering, M. J., and McElree, B. (2002). Coercion in Sentence Processing: Evidence from Eye-Movements and Self-Paced Reading. *Journal of Memory and Language*, 47(4): 530-547.
Zarcone, A. and Pado, S. (2011). Generalized Event Knowledge in Logical Metonymy Resolution. *In Proceedings of the Annual Meeting of the Cognitive Science Society*, 33.
Zarcone, A., Utt, J., and Padó, S. (2012). Modeling Covert Event Retrieval in Logical Metonymy: Probabilistic and Distributional Accounts. In R. Levy and D. Reitter (eds.) *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 70-79, Montréal, Canada, June 7, 2012.
Zarcone, A., Lenci, A., Pado, S., and Utt, J. (2013) Fitting, Not Clashing! A Distributional Semantic Model of Logical Metonymy. *In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Short Papers*.
Zarcone, A., Padó, S., and Lenci, A. (2014). Logical Metonymy Resolution in a Words-as-Cues Framework: Evidence from Self-paced Reading and Probe Recognition. *Cognitive Science*, 38(5):973-996.
Zarcone, A., McRae, K., Lenci, A. and Pado, S. (2017). Complement Coercion: The Joint Effects of Type and Typicality. *Front. Psychol*. 8:1987.

Appendix: Fully Described Prompts

Question: What are the lexical semantic constraints for a complement of the transitive verb 'begin'?
Answer: The lexical semantic constraints for a complement of the transitive verb 'begin' are related to the nature of the verb itself. 'Begin' is an aspectual verb, which forms a subset of the class of 'coercion verbs' implicated in the phenomenon of complement coercion. Aspectual verbs like 'begin' have been assumed to <u>select exclusively for eventive complements</u>, such as in the sentence 'John began reading the book'.

Prompt 1 Complement selection constraints of the aspectual verb 'begin'

Question: Please list the most common verbs that can be inserted into "John began to _____ the book." in order of frequency.
Answer: … 1. Read, 2. Write, 3. Study, 4. Open, 5. Close, 6. Finish, 7. Review, 8. Mark, 9. Cover, 10. Illustrate, 11. Edit, 12. Publish, 13. Sell, 14. Buy, 15. Bind, 16. Sign, 17. Dedicate, 18. Translate, 19. Print, 20. Annotate.

Prompt 2 The most common twenty transitive verbs that co-occur with 'begin' and 'the book'

Question: Please list the 20 transitive verbs that cannot be inserted into 'John began to ____ the book.' with the reason.
Answer: 1. Swim, 2. Jump, 3. Run, 4. Fly, 5. Breathe, 6. Eat, 7. Drink, 8. Sleep, 9. Dance, 10. Sing, 11. Laugh, 12. Cry, 13. Sneeze, 14. Walk, 15. Drive, 16. Climb, 17. Smile, 18. Frown, 19. Blink, 20. Sweat
These verbs are awkward in this context because they denote <u>actions that are not typically associated with a book or reading</u>. Remember, language use can vary greatly depending on context, and what seems awkward in one situation might be perfectly natural in another.

Prompt 3 The most improbable transitive verbs that co-occur with "begin and "the book"

Sentence: John began to obtain the book.
Question: Is the sentence semantically acceptable?
Answer: <u>Yes</u>, the sentence "John began to obtain the book." is semantically acceptable. The verb "obtain" means to come into possession of something, so this sentence suggests that <u>John started the process of getting the book</u>. This could be through buying, borrowing, or receiving it as a gift, among other possibilities. The sentence does not specify whether John completed the action of obtaining the book, only that he started it. The appropriateness of its use would depend on the specific context.

Prompt 4 Question regarding the possibility of co-occurrence of 'obtain' and 'begin'

Sentence: John obtained the book <u>for three hours</u>.
Question: Is the sentence semantically acceptable?
Answer: <u>No</u>, the sentence "John obtained the book for 3 hours" is <u>not semantically acceptable</u>. The verb "obtain" implies a change of possession, which is typically an instantaneous action rather than one that extends over a period of time. Therefore, it's unusual to use "for three hours" with "obtain". A more appropriate sentence might be

"John read the book for 3 hours." or "John searched for the book for 3 hours.", depending on the intended meaning. Remember, the appropriateness of a verb can depend on the specific context.

Prompt 5 Is Copilot knowledgeable about the lexical semantic properties of achievement verbs?

Question: What is the semantic type of the English noun "book"?

Answer: The semantic type of the English noun "book" is associated with the category of "things" or "objects".

Prompt 6 Semantic Type of the English Noun "book"

Sentence: John began the book.
Question: What is the meaning of the sentence?

More balanced mode answer: The sentence "John began the book" typically means that John started reading a book or started writing a book. The word "began" implies the initiation of an action, and in this context, it refers to the start of John's engagement with the book.
More precise mode answer: … when no specific context is provided, the most common understanding is that John started reading or writing the book. The exact meaning depends on the context.
More creative mode answer: The sentence 'John began the book' conveys the idea that John initiated the act of starting to read or write the book. The exact interpretation depends on the context.

Prompt 17 Interpretation in the More Balanced session

Sentence: John began the book. **Question**: What is the most common interpretation of the sentence when no specific context is given in English?

Answer: The most common interpretation of the sentence "John began the book" when no specific context is given in English is that John started reading a book. However, it could also mean that John started writing a book, depending on the context. Without additional information, it's difficult to determine the exact meaning. The verb "began" is quite versatile and can be used in various contexts. It's always best to consider additional context when interpreting sentences.

Prompt 8 The most common interpretation of the Aspectual Coercion sentence when no specific context is provided

Sentence: John began the book.
Question: Please list the 20 possible interpretations of the sentence.

Answer: Sure, here are 20 possible interpretations of the sentence "John began the book": 1. John started reading a book. 2. John started writing a book. 3. John opened the book to start reading. 4. John started studying the book for academic purposes. 5. John began narrating the book to someone else. 6. John started translating the book into another language. 7. John started illustrating a book. 8. John began binding the book. 9. John started selling the book. 10. John began a book club focused on the book. 11. John started a journey that the book is about. 12. John began to critique the book. 13. John started to adapt the book into a movie or play. 14. John began to memorize the content of the book. 15. John started to audio-record the book. 16. John began to teach the content of the book. 17. John started to convert the book into braille. 18. John began to digitize the book. 19. John started to restore the old book. 20. John began to research about the book's background. Please note that the exact interpretation would depend on the context in which the sentence is used.

Prompt 9 Twenty possible interpretations of the sentence

Question: What were the criteria for selecting the 20 interpretations?
Answer: Here are the criteria I used:

1. **Different Actions**: The verb "began" can imply various actions such as reading, writing, translating, selling, etc. Each action provides a different interpretation of the sentence.
2. **Different Roles**: The subject "John" can take on different roles in relation to the book – reader, writer, seller, teacher, etc. Each role provides a different interpretation.
3. **Different Contexts**: The context in which the sentence is used can drastically change its meaning. For example, in a publishing context, "John began the book" could mean he started writing it. In a bookstore, it could mean he started selling it.
4. **Feasibility**: The interpretations are plausible scenarios that could realistically occur in everyday life.
5. **Variety**: To provide a wide range of interpretations, I considered as many different actions and roles as possible.

Remember, the exact interpretation of a sentence often depends on the context in which it's used. Without additional context, multiple interpretations could be valid.

Prompt 10 Copilot's criteria for selecting twenty possible interpretations

Context: John is an editor who works for a publishing company. **Sentence**: John began the book. **Question**: What is the most appropriate interpretation of the sentence when the context is given?
Answer: Given the context that John is an editor working for a publishing company, the most appropriate interpretation of the sentence "John began the book" would be that John started editing the book. This interpretation assumes that "the book" refers to a manuscript or a draft that John is

working on as part of his job as an editor. Remember, the exact interpretation of a sentence often depends on the context in which it's used. Without additional context, multiple interpretations could be valid.

Prompt 11 Interpretation of the sentence when a specific context is provided

Question: Under the context, is the only interpretation you gave possible? Please give me other possible 10 interpretations of the sentence when the context is given.
Answer: Given the context that John is an editor working for a publishing company, here are 10 other possible interpretations of the sentence "John began the book":
1. John started to review the book for potential publication. 2. John began to format the book for printing. 3. John started to plan the layout of the book. 4. John began to proofread the book. 5. John started to assign the book to a specific genre or category. 6. John began to write a summary or blurb for the book. 7. John started to coordinate with the author about revisions for the book. 8. John began to plan the marketing strategy for the book. 9. John started to discuss the book with other editors or staff. 10. John began to prepare the book for digital publication.

Prompt 12 Ten other possible interpretations

Question: Is it impossible to interpret the sentence as 'John began reading the book' under the context?
Answer: No, it's not impossible. Even though John is an editor and his job involves tasks like reviewing, proofreading, and formatting, it's still plausible that "John began the book" could mean "John began reading the book". Perhaps he's reading it as part of his editing process, or maybe he's reading it for pleasure. Context is key, and without more specific information, multiple interpretations can still be valid. Remember, the exact interpretation of a sentence often depends on the context in which it's used. Without additional context, multiple interpretations could be valid.

Prompt 13 Is reading interpretation possible under the context in Prompt 12?

Sentence: John began the book. Question: Please list the 20 impossible interpretations of the sentence.

Answer: **1.** John started to eat the book. **2.** John began to use the book as a spaceship. **3.** John started to wear the book as a hat. **4.** John began to play the book like a guitar. **5.** John started to drink the book. **6.** John began to use the book as a time machine. **7.** John started to fly using the book. **8.** John began to use the book as a telescope. **9.** John started to use the book as a pair of shoes. **10.** John began to use the book as a magic wand. **11.** John started to use the book as a football. **12.** John began to use the book as a swimming pool. **13.** John started to use the book as a parachute. **14.** John began to use the book as a cooking plan. **15.** John started to use the book as a pillow. **16.** John began to use the book as a car. **17.** John started to use the book as a boat. **18.** John began to use the book as a paintbrush. **19.** John started to use the book as a basketball hoop. **20.** John began to use the book as a refrigerator.

Prompt 18 Impossible interpretations

Context: John's father gave a chocolate box in which there are various shapes of chocolates. One of them is a book-shape. **Sentence**: John began the book. **Question**: What is a common interpretation of the sentence when the context is given?
Answer: Given the context, a common interpretation of the sentence could be that John started eating the chocolate that is shaped like a book. This interpretation assumes that "the book" refers to the book-shaped chocolate in the context provided. Remember, the exact interpretation of a sentence often depends on the context in which it's used. Without additional context, multiple interpretations could be valid.

Prompt 15 Common interpretation of the sentence under a specific situational context

Words = ['begin', 'finish'] **Sentence**: John _____ the book. **Question**: For each word in words, please make a sentence by inserting each word into ____. And then please list 5 possible interpretations of each sentence in order of frequency.
Answer: Here are the possible interpretations for each word:
1. John began the book.
- John started writing a novel, perhaps sitting at his desk with a blank page before him.
- John initiated the process of creating a literary work, whether fiction or non-fiction.
- John embarked on the journey of authorship, with ideas swirling in his mind.
- John opened a new chapter in his creative endeavors.
2. John finished the book.
- John completed reading a captivating novel, turning the final page with satisfaction.
- John concluded writing a manuscript, typing the last sentence and feeling a sense of accomplishment.
- John reached the end of a literary work, whether it was a short story or an epic saga.
- John wrapped up the book project, editing and revising until it was polished.
- John closed the cover of a well-loved book, reflecting on the journey it took him on.
Remember, context matters, and the intended meaning depends on the surrounding information!

Prompt 16 Building a lexical resource about Aspectual Coercion

Can GPT-4 Recover Latent Semantic Relational Information from Word Associations? A Detailed Analysis of Agreement with Human-annotated Semantic Ontologies.

Simon De Deyne[1] **Chunhua Liu**[2] **Lea Frermann**[2]
[1]Complex Human Data Hub, [2]School of Computing and Information Systems
The University of Melbourne
{simon.dedeyne, chunhua.liu1, lea.frermann}@unimelb.edu.au

Abstract

Word associations, i.e., spontaneous responses to a cue word, provide not only a window into the human mental lexicon but have also been shown to be a repository of common-sense knowledge and can underpin efforts in lexicography and the construction of dictionaries. Especially the latter tasks require knowledge about the relations underlying the associations (e.g., Taxonomic vs. Situational); however, to date, there is neither an established ontology of relations nor an effective labelling paradigm. Here, we test GPT-4's ability to infer semantic relations for human-produced word associations. We use four human-labelled data sets of word associations and semantic features, with differing relation inventories and various levels of annotator agreement. We directly prompt GPT-4 with detailed relation definitions without further fine-tuning or training. Our results show that while GPT-4 provided a good account of higher-level classifications (e.g., Taxonomic vs Situational), prompting instructions alone cannot obtain similar performance for detailed classifications (e.g., superordinate, subordinate or coordinate relations) despite high agreement among human annotators. This suggests that latent relations can at least be partially recovered from word associations and highlights ways in which LLMs could be improved and human annotation protocols could adapted to reduce coding ambiguity.

Keywords: Large language models, semantic relations, word associations

1. Introduction

The word association test (WAT) provides important information about the organisation of the mental lexicon. In a typical study, participants are presented with a cue word (e.g., *dog*) and produce the first word(s) that come to mind (e.g., *cat* or *bark*). This procedure is often referred to as *free* word association as participants are not restricted in their responses, making it one of the most general methods to obtain subjective behavioural estimates of word meaning (Deese, 1965).

In recent years, online crowd-sourcing approaches such as the Small World of Words project have demonstrated that this approach is highly scaleable, with several datasets including millions of responses published in Dutch, English, Spanish and Chinese (De Deyne et al., 2019). As such, word associations provide valuable resources for the fields of lexicography and semantic typology, which study the availability and organization of senses and meaning within and across languages.

A common type of analysis of these data involves classifying responses according to a semantic ontology that covers taxonomic (*dog* – *cat*), concept properties (*dog* – *tail*), situational properties (*dog* – *park*) or introspective properties (*dog* – *friend*). This is of interest to cognitive science, where these classifications can shed light on the nature of our mental representation and the time course over which this information becomes available (Fitzpatrick and Thwaites, 2020), (Garrard et al., 2001), metaphor comprehension and analogies (Lu et al., 2022).

Word associations have also shown promise as a tool to derive common sense knowledge (Liu et al., 2021). In this respect, recent work suggests that they could fill the gaps in other lexical knowledge graphs. While word associations do not capture the depth of other approaches (e.g., the number of senses of a word), they do capture frequent senses and measure what aspects of meaning are dominant among a community of speakers. Importantly, word associations are informed not only by our linguistic environment but encode extra-linguistic experiential information as well that is difficult to reveal by only studying how words co-occur in language (Fitzpatrick and Thwaites, 2020). Various supervised and unsupervised approaches to predict associations from text had correspondingly mixed success (Griffiths et al., 2007; Cattle and Ma, 2017; Liu et al., 2022).

Large language models (LLMs) like GPT-4 (Achiam et al., 2023) have shown unprecedented abilities not only to generate naturalistic text but also to support complex data annotation (Gilardi et al., 2023) and annotation of single words or word pairs for comparison with human similarity judgments, induction and lexical ratings (e.g., concreteness) (Han et al., 2024; Marjieh et al., 2023; Trott, 2023).

The Workshop on Cognitive Aspects of the Lexicon, pages 68–78
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Here we test the ability of GPT-4 (Achiam et al., 2023), a state-of-the-art LLM, to recover semantic relations for human-produced word associations. This is of interest for three reasons. First, this new generation of models, with the capacity to encode long prompts, does not have the same working memory constraints human annotators have when confronted with extensive fine-grained semantic ontologies. Second, we extend a line of work that assesses the utility of LLMs as cognitive models to the task of semantic relation labelling. Third, from a practical perspective, a model that can automatically predict semantic relations can support the construction or augmentation of lexical or commonsense databases.

A highly influential ontology in the cognitive sciences is the Wu-Barsalou (WB) ontology (Wu and Barsalou, 2009). The WB scheme is hierarchically organised and consists of four major relation types, which we will refer to as Level 1: Taxonomic relations, Entity/concept properties, Situational properties and Introspective properties. More detailed Level 2 distinctions are nested within each relation class (e.g., Subordinate, Coordinate are properties nested under Taxonomic relations). While the WB ontology was initially developed to investigate grounding in semantic representations, it has since been applied broadly across many property listing tasks (PLT, see Bolognesi et al., 2017, for an overview) and was recently adapted to the WAT (Liu et al., 2022; Chen et al., 2024). The scheme has been adapted over the years to suit the needs of individual researchers. However, these changes tend to be minor simplifications of Level 2 distinctions such as grouping Buildings and Location or Subordinates and Individuals (see Bolognesi et al., 2017; Liu et al., 2022).

In contrast to the WAT, the PLT is often assumed to be less ambiguous and consequently easier to annotate because the properties can be phrases (e.g., *dog - is a kind of animal*) that can be easily mapped onto the ontology (Superordinate). However, an inspection of existing semantic feature generation studies suggests that features are often coded by the annotators as a single-word response (e.g., *zebra – horse*), similar to a word association. In the PLT of Vivas et al. (2021), for example, 18% of features consisted of a single words, whereas in Bolognesi et al. (2017), 92% of features consisted of a single word. Consequently, presumed ambiguity is not limited to word association per se but is also highly prevalent in semantic feature norms.

1.1. Current work

This study will use previously annotated datasets for word associations and semantic features. The latter are included as contrast cases that allow us to contextualise our findings, as the semantic relation is often included in the participant response. We focus primarily on the Wu-Barsalou semantic ontology (WB), which is widely used in cognitive psychology and GPT-4 as a state-of-the-art (SOTA) LLM. To the best of our knowledge, we are the first to use GPT-4 for the task of relation labelling, despite its remarkable performance for related tasks where limited context is available (e.g., pairwise similarity judgments). Focusing on a single model provides us with an opportunity to analyse (mis)classification and inconsistencies across the different datasets and levels in the label hierarchy. In sum, we address the following research questions:

- To what degree can latent semantic relations be recovered in SOTA LLMs?
- How does performance vary for broad vs fine-grained relation labels?
- How does the nature of the task (WAT vs PLT) affect the results?
- What are the most common confusions, and to what degree do these reflect limitations of the model or inherent ambiguity due to word association data or existing coding schemes?

2. Methods

We introduce the primary relation ontology, which researchers have adopted for classifying word associations, the datasets, and the LLM that will be used in current work.

2.1. Datasets

The current study includes four recent datasets. Studies were included according to the following criteria: 1) the use of the WB scheme (or a close derivative) for the relation annotation; 2) including a large number of concepts; 3) the availability of English translations in the published data for non-English datasets; and 4) the use of multiple annotators with the inter-rater agreement information included in the original study. All data sets share the same four Level 1 relations (Taxonomic, Entity, Situational, and Introspective) but differ in their Level 2 labels. See more details of the labels along with other dataset statistics in Table 1.

2.1.1. Bolognesi-2017

The PLT dataset in Bolognesi et al. (2017) consists of English concept-feature pairs that were carefully annotated through an ontology that resembles a decision tree. This relation ontology has been updated from the WB ontology to accommodate both concrete and abstract concepts effectively. The resulting dataset had a high inter-rater agreement

Bolognesi-2017: #C = 180, #(C,R) = 1919, #L2 = 20

TAXONOMIC RELATION (T): Synonyms, description and linguistic clues (syn), Antonyms (ant), Superordinates (sup), Subordinates and instances (sub), Coordinate (coor)

ENTITY PROPERTY (E): Perceptual properties (perc), Non-perceptual properties (sys), Components, materials and substances (comp), Larger wholes, thematic larger wholes, and disciplines (whol), Entity behaviors (beh)

SITUATIONAL PROPERTY (S): Objects (obj), Participants (par), Actions (act), Properties of contextual entities (other), Function (fun), Locations, containers, and buildings (loc), Time and events (time)

INTROSPECTIVE PROPERTY (I): Evaluations (eval), Emotions (emo), Contingencies and complex cognitive operations (cont)

Vivas-2022 #C= 400, #(C,R) = 2669, #L1 = 33

TAXONOMIC RELATION (T): Synonym (syn), *Ontological category (ont)*, Superordinate (super), Coordinate (coord), Subordinate (subord)

E: *External component (excomp)*, *Internal component (incomp)*, *External surface property (exsurf)*, *Internal surface property (insurf)*, Substance/Material (mat), *Spatial relation (spat)*, Systemic property (sys), Larger whole (whole), Entity behavior (beh), Abstract entity property (abstr)

SITUATIONAL PROPERTY (S): Person (person), Living thing (living), Object (object), Social organization (socorg), *Social artifact (socart)*, *Building (build)*, Location (loc), Spatial relation (spat), Time (time), Action (action), Event (event), Function (func), Physical state (physt), Social state (socst)

INTROSPECTIVE PROPERTY (I): Affect/emotion (emot), Evaluation (eval), *Representational state (rep)*, *Cognitive operation (cogop)*, Contingency (contin), Negation (neg)

Chen-2024 #C = 505, #(C,R) = 2292, #L2 = 21

TAXONOMIC RELATION (T): Synonym (syn), Superordinate (super), Coordinate (coord), Subordinate (sub), Antonym (ant)

ENTITY PROPERTY (E): Components/Material/Substance (comp), Whole (E-whole), Entity property (prop), Entity behavior (beh), Typical state (state)

SITUATIONAL PROPERTY (S): Function (function), Location/Container/Building (loc), Object (obj), Action (action), Agent (agent), Time/Events (time), Contextual entity property (context), Situational state of target (targetstate)

INTROSPECTIVE PROPERTY (I): Evaluation (eval), Emotion (emo), Contingencies and complex cognitive operations (contin)

Liu-2022 #C = 340, #(C,R) = 476, #L2 = 15

TAXONOMIC RELATION (T): Synonym (syn), Antonym (ant), *Category-Exemplar-Pairs (cat)*, Members-of-same-Category (coord)

ENTITY PROPERTY (E): PartOf (part), Material-MadeOf (mat), *HasProperty (prop)*

SITUATIONAL PROPERTY (S): Time (time), Location (loc), Function (func), *Has-Prerequisite (preq)*, *Result-In (result)*, Action (action), *Thematic (them)*

INTROSPECTIVE PROPERTY (I): Emotion-Evaluation (emo)

Table 1: Summary of datasets. #C denotes the number of unique cues, #(C,R) denotes the number of unique cue-response pairs, #L2 denotes the number of Level 2 relations. The dataset-specific L2 labels are in italics.

with Cohen's κ = .886 for the Level 1 distinctions, and κ = .866 for the Level 2 distinctions.

2.1.2. Vivas-2022

The Vivas et al. (2021) Features PLT dataset consisted of noun-feature pairs collected from Spanish speakers across a range of concrete semantic domains. The reported inter-rater agreement measured as Krippendorff's α was high: .78 for novice coders and .86 for trained coders (Vivas et al., 2021). The ontology closely followed the original WB scheme. In the current analyses, we did not include additional quantifier codes and two codes that were not used by any of the annotators (C-INDIV and S-MANNER). A separate set of Meta-codes (e.g., hesitations, repetition, comments) was also not included in the current results.

A second dataset, *Vivas-2022 Asso*, was derived by extracting a key word (e.g., *zebra*, *music instrument*). This way, additional relational cues such as <is a> were removed, allowing us to define a baseline to determine how these relation indicators reduce ambiguity when annotating PLT data.[1]

2.1.3. Chen-2024

The Chen et al. (2024) WAT data consists of a semantic ontology derived from the WB ontology. The cues and responses were derived from the English Small World of Words project (De Deyne et al., 2019). The stimuli comprised 507 nouns (ranging in concreteness) and their top 5 associative responses. All cue-response pairs were coded by two trained coders for broad (Level 1) and fine-grained (Level 2) distinctions. For this study, we only used the Taxonomic, Entity, Situational and Introspective Level 1 properties (see Table 1 for a list of included Level 2 properties). We did not include form position-based properties since these could also be estimated from word co-occurrence data directly and overlap significantly with semantic properties and also omitted meta codes (e.g., erroneous responses) similar to the approach for the Vivas-2022 dataset. The inter-rater agreement, measured as Cohen's κ, was high, .81, for both Level 1 and Level 2 relations.

2.1.4. Liu-2022

The Word Association Explanation database (WAX) (Liu et al., 2022) includes word associations for a total of 15K different English cue-response pairs. A subset of 520 pairs was annotated with semantic relations. Human coders were recruited through Amazon Mechanical Turk. The ontology represents a simplification of the WB ontology, focusing on the main types across all of the four major Level 1 distinctions. The Level 2 properties also included a few additional relations from ConceptNet (Speer et al., 2017) for event-related associations (e.g., Has-Prequisite, Result-In). The pairwise annotator agreement was moderate, Cohen's κ = 0.42.

Like Chen-2024, we did not include linguistic and form-based responses (e.g., Sound Similarity, Common Phrases). An unspecified category (None-of-the-above) was also removed. Finally, note that Emotion-Evaluation were originally grouped under Concept/entity properties. For reasons of comparability, we decided to move this property to a separate Level 1 Introspective properties section consistent with the other datasets.

[1] The Bolognesi-2017 dataset consisted mainly of single words, and so this procedure was not applied.

2.2. SOTA LLM Model

We used GPT-4 (Achiam et al., 2023) through the OpenAI API and specified model version *gpt-4-0613*. Across all studies, the temperature was set to 0, and no optional system prompts were provided. Cue-response pairs were randomized and split into batches of 100 items before being concatenated to the instruction prompt.

2.2.1. Prompting

All prompts followed the same structure at the start and end but differed in terms of the definitions and examples, which were taken from the original articles. All materials and prompts are available in the original articles and online repository.[2] The default prompt was as follows:

You will be presented with a list of word pairs consisting of an associated cue and an associated target word separated by ' – '.

You are asked to choose a code with square brackets [] that best describes the semantic relation between the cue and the target word. Each code refers to a specific semantic relation that refers to Taxonomic properties, Concept properties, Situation properties, or Introspective properties.

We will now provide you with a definition and examples for each of these, which you will carefully consider when choosing one of the codes.

{Relation taxonomy with definitions and examples.}

Remember to only choose from the above codes between square brackets. Do not further elaborate on your response. Format your response as follows cue — target: code.

List:

{List of 100 cue association pairs: }

For the Bolognesi-2017, Vivas-2022 Features and Vivas-2022 Association, the first sentence was replaced by "You will be presented with a list of word pairs consisting of a cue and a semantic feature separated by '–'. ". Finally, consistent with the instructions in (Vivas et al., 2021), we added "In these examples, the relation signified by the semantic feature is higlighted by using capitalized letters." after the third sentence ("Each code refers...").

3. Results

3.1. Response preprocessing

All responses were provided in the cue — target: code format consistent with the instructions, which

[2] Materials, instruction prompts with definitions and examples as well as the analysis scripts are available at https://github.com/SimonDeDeyne/lrec2024

means no further manual extraction was required. On a very small number of occasions, erroneous codes (i.e., codes not in the instructions were returned). These were subsequently removed.

3.2. Classification

For each of the datasets, we calculated accuracy, precision, recall, macro-F scores, and Kappa inter-rater reliability at both Level 1 (broad) and Level 2 (detailed). Results are presented for a cue-response type-based classification and a token-based classification, where the latter is weighted by the number of times participants generated a particular response. This provides information that is more useful for real-world settings where only a subset of cue-response pairs might be inspected, which means that accurate relation labels are especially important for the most frequent responses.

Since the response classes (i.e., relation labels) are unbalanced, classification metrics were weighted by prevalence (between 0 and 1) before averaging over classes. With these balanced scores, the role of relatively infrequent classes, such as Introspective properties, which were rare, was proportionate when averaging all four Level 1 classes.

Unlike other datasets, the Chen-2024 included the codes for two individual annotators, A and B. Unless stated otherwise, we also provide the results for the LLM and individual coder agreement.

The results are shown in Table 2 and Table 3. The last three columns show the baseline performance using the majority class (MC), a score where only the majority relation class was considered and which is contrasted with accuracy (see Table 2 and 3). In all cases, the accuracy rate significantly differed from the MC baseline.

3.2.1. Type-based results

The results in Table 2 show high values across all metrics for the feature datasets (Bolognesi et al., 2017; Vivas et al., 2021) at Level 1 and moderate results at Level 2 of the ontology. The results of deriving pseudo-associations after censoring semantic relations from features for the Vivas-2022 Asso dataset had a negligible effect at Level 1 and only a minor drop in performance at Level 2. The results for the two word associations sets (Chen et al., 2024; Liu et al., 2022), were somewhat lower, with good results at Level 1 and moderate to low results at Level 2. The agreement between the LLM predictions and individual coders for Chen-2024 was highly consistent for A and B, with slightly better results for annotator A. However, comparing the scores with those obtained by directly comparing annotators A and B (see Chen AB in Tables2) suggests some room for further improvement when benchmarked against trained human annotators, and this is notably the case for the Level 2 Ontology annotations.

3.2.2. Token-based results

To calculate performance that considers how frequently the responses are generated, weighted results were calculated on the raw data before tabulation. Doing so provides an estimate of classification performance that is more relevant for applications and also allows us to determine whether infrequent responses are inherently more difficult to classify. Consistent with this, Table 3 shows results that are largely consistent with Table 2, albeit slightly higher. The only exception to this pattern was the Liu-2022 dataset, where the difference was less pronounced, which is likely to reflect the relatively small range of frequency given the limited number of cue presentations in this dataset. Similar to the type-based results, LLM prediction performed comparably across annotators in the Chen-2024 dataset but was still lower compared to the results when comparing two trained human annotators. For simplicity, we will only consider the results for Coder A and the remaining analysis.

3.3. Error analysis

Token-based confusion matrices for the Level 1 distinctions are plotted in Figure 1. Each cell encodes the proportion of cross-classifications and supplements Table 3. The main focus is on the entries on off-diagonal elements, which indicate systematic differences between human coders and the model classification. Note that the values do not have to be symmetric. For example, in the Bolognesi dataset, 2% of the responses humans consider introspective were coded taxonomic. Vice versa, only 1% of the responses humans code as taxonomic are labeled as introspective.

Consistent confusion was present in the Bolognesi data for Introspective properties across most other L1 relations. Closer inspection showed that many of the pairs were coded as “Contingencies and complex cognitive operations”. Relative large proportions of these confusions were also found for Taxonomic vs Entity properties (0.06 for Chen-2024). In addition, Taxonomic and Entity properties were also frequently confused in the Liu-2022 dataset (0.09). Insightful examples include *genius* – *brilliant*, which human annotators code as an entity property, but GPT-4 considers a synonym. This highlights the fact that the model does not capture a human noun-bias, which is typical in association data where words are ambiguous in terms of part of speech. Another example is *lonely* – *depressed*, which was also considered a synonym but coded as a “Result-In” feature by the annotators. More

	Level 1 Ontology					Level 2 Ontology				
	MC	acc.	κ	prec.	F1	MC	acc.	κ	prec.	F1
Bolognesi	0.388	0.717	0.609	0.765	0.730	0.116	0.542	0.510	0.647	0.564
Vivas Feat	0.438	0.846	0.768	0.867	0.851	0.207	0.614	0.584	0.720	0.618
Vivas Asso	0.401	0.846	0.769	0.852	0.847	0.129	0.599	0.571	0.623	0.577
Chen A	0.419	0.763	0.653	0.769	0.764	0.194	0.523	0.487	0.615	0.536
Chen B	0.419	0.713	0.584	0.738	0.716	0.194	0.492	0.454	0.607	0.499
Liu	0.420	0.718	0.562	0.758	0.727	0.282	0.464	0.399	0.535	0.471
IAA (Chen AB)	0.357	0.880	0.825	0.888	0.879	0.127	0.808	0.794	0.822	0.809

[a] All accuracy vs MC comparisons were significant, $p < .001$.
[b] Recall is identical to accuracy after prevalence weighting.

Table 2: Type-based classification results (acc. = accuracy, MC = Majority Class, κ, prec. = precision, F1) for the Level 1 (left) and 2 (right) ontologies across semantic feature (Bolognesi-2017, Vivas-2022 Feat) and word association (Vivas-2022 Asso, Chen-2024, Liu-2022) datasets. We list agreement with GPT-4 for individual annotators (Chen A and B), alongside inter-annotator scores for annotators A and B of Chen-2024 (IAA ChenAB).

	Level 1 Ontology					Level 2 Ontology				
	MC	acc.	κ	prec.	F1	MC	acc.	κ	prec.	F1
Bolognesi	0.389	0.733	0.630	0.778	0.746	0.126	0.566	0.535	0.664	0.586
Vivas Feat	0.413	0.865	0.797	0.882	0.869	0.225	0.622	0.591	0.726	0.623
Vivas Asso	0.409	0.860	0.791	0.865	0.861	0.138	0.594	0.565	0.606	0.565
Chen A	0.422	0.780	0.676	0.787	0.781	0.194	0.540	0.502	0.624	0.549
Chen B	0.422	0.730	0.606	0.755	0.732	0.194	0.507	0.468	0.614	0.509
Liu	0.407	0.723	0.574	0.762	0.729	0.272	0.489	0.425	0.544	0.491
IAA (Chen AB)	0.344	0.884	0.830	0.893	0.884	0.139	0.816	0.801	0.830	0.817

[a] All accuracy vs MC comparisons were significant, $p < .001$.
[b] Recall is identical to accuracy after prevalence weighting.

Table 3: Token-based classification results (acc. = accuracy, MC = Majority Class, κ, prec. = precision, F1) for the Level 1 (left) and 2 (right) ontologies across semantic feature (Bolognesi-2017, Vivas-2022 Feat) and word association (Vivas-2022 Asso, Chen-2024, Liu-2022) datasets. We list agreement with GPT-4 for individual annotators (Chen A and B), alongside inter-annotator scores for annotators A and B of Chen-2024 (IAA ChenAB).

.

generally, GPT-4 tends to be biased towards taxonomic responding, which is not always incorrect, but highlights the fact that relation types are not mutually exclusive.

The remainder of the error analysis at the detailed Level 2 will primarily focus on the word association datasets (Chen-2024 and Liu-2022). The micro-level confusion matrix for the Chen-2024 dataset shown in Figure 2 indicates a combination of confusion within and between macro-categories. As shown in the upper left corner, the LLM struggles to distinguish between different types within the Level 1 Taxonomy group, favoring Synonymy over Coordinate, Superordinate and Subordinate relations. The LLM also confuses Synonyms with Entity properties and Entity components. Examples of Entity properties include confusion where Large Wholes are confused with Situated-objects and Entity components. Among Situation properties, functions and actions are also frequently confused.

As can be seen from the large proportion of highlighted off-diagonal elements in Figure 3, confusion is spread across all four major semantic relation categories. It is seemingly lower for Taxonomic categories, although it should be noted that the Liu-2022 ontology does not distinguish between Subordinates and Superordinate relations, which might skew the comparison with Chen-2024. Beyond Level 1 confusion in Figure 1, Figure 3 shows that different types of Situation properties are not clearly distinguished.

To illustrate, Figure 3 shows that situational actions (S-act) and thematic relations (S-them) are easily confused. This is also an interesting case. The former is defined in the instructions as "An

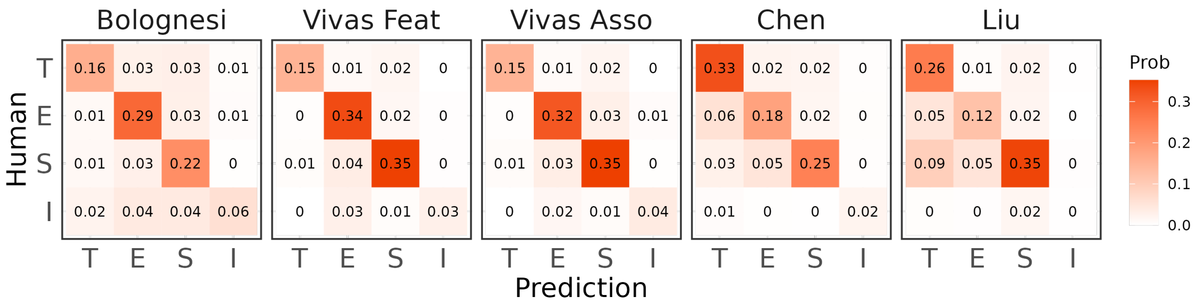


Figure 1: Confusion matrices for macro-level distinctions across five datasets (Properties: T = Taxonomic, E = Entity/Concept, S = Situation, I = Introspective).

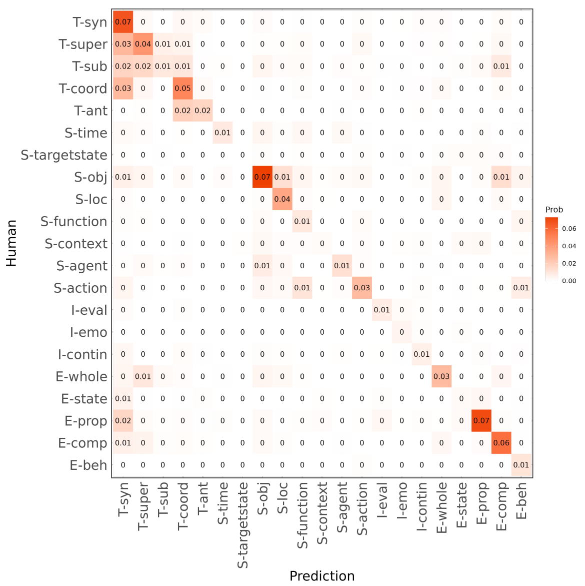


Figure 2: Confusion matrix for the Chen-2024 dataset showing a cross-tabulation of proportions for GPT-4 on the x-axis and human (coder A) reference classification on the y-axis.

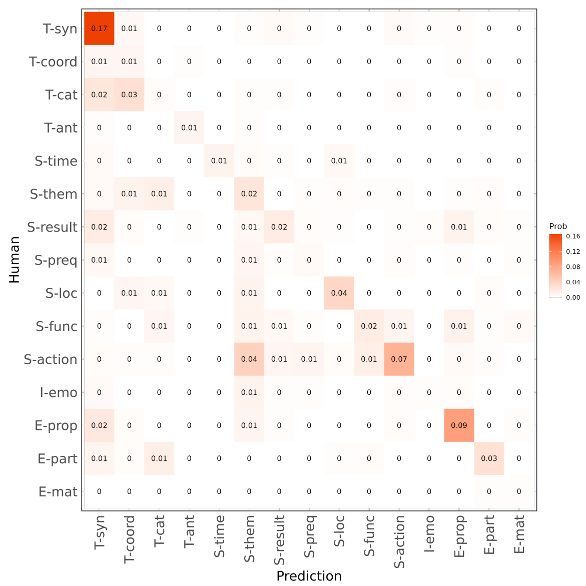


Figure 3: Confusion matrix for the Liu-2022 dataset analogues to Figure 2.

action that a participant (could be the cue, association or others) performs in a situation., whereas thematic relations are defined as “Cue and association participate in a common event or scenario. None of the other situational properties applies.”. Examples of misclassified actions include *dollar – earn* and *running – race*. Still, other pairs like *tactful – conversation* that were labelled S-act by humans but S-them by the LLM illustrate the presence of false negatives (and potential limitations of the original ontology) as well.

4. Discussion

In this study, we investigated to what degree GPT-4 can recover the latent semantic relations in word association tasks. While the findings pertain to datasets with different stimuli and different variants of the WB ontology, the overall pattern of results was consistent. First, our results across two word association datasets using GPT-4 showed good performance in making broad distinctions regarding Taxonomic, Entity, Situation and Introspective properties. More fine-grained distinctions were predicted only partially, despite relatively high levels of human inter-annotator agreement. This suggests room for further improvements, although procedural aspects such as calibration or consensus coding, which are commonly employed in human annotations, make this comparison less straightforward.

Second, a comparison with human data derived from the Property Listing Task showed high performance in capturing broad distinctions and good to moderate performance in making fine-grained distinctions. Moreover, this performance was not entirely driven by the fact that the responses in the property listing task are less ambiguous. Even when disambiguation information in the form of explicit indicators was removed, and only a single word was retained, performance was similar. Furthermore, the performance for word associations was on par when compared to the Bolognesi-2017 dataset that covered a more challenging set of cues by including many abstract concepts and single-word responses.

4.1. Comparison with previous work

As far as we know, previous work that has used LLMs to predict semantic relations using the WB taxonomy is limited. One exception is the work by Liu et al. (2022), in which a subset of training relations was used to fine-tune BERT (Devlin et al., 2019) and BART-Large (Lewis et al., 2020) to predict performance among a test set of the Liu-2022 relations. We investigated how BART, the best-performing model, compared with GPT-4 for 88 unique cue-response pairs shared among both datasets.

Across all analyses, the results showed that the GPT-4 outperformed BART. Illustrating this with the token-based analysis, the results for the L1 level were GPT-4: accuracy = 0.732, κ = 0.593, precision: 0.781, F1 = 0.733; and BART: accuracy = 0.653, κ = 0.479, precision: 0.645, F1 = 0.641. At the more detailed L2 level, we obtained for GPT-4: accuracy = 0.521, κ = 0.455, precision: 0.621, F1 = 0.535; and BART: accuracy = 0.493, κ = 0.431, precision: 0.540, F1 = 0.491. Interestingly, when comparing both types of LLMs, their mutual agreement was higher than that obtained against human annotators. For the L1 level: accuracy = 0.756, κ = 0.623, precision: 0.775, F1 = 0.742 and for the L2 level: accuracy = 0.577, κ = 0.523, precision: 0.616, F1 = 0.562. This suggests that the relations predicted by different types of language models might have more in common with human annotators. That said, given the small number of pairs in this comparison, more work is needed before strong conclusions can be drawn.

4.2. How ambiguous are word associations?

One way of determining to what degree word associations can be annotated is by comparing the relative performance for agreement among human annotators and LLM predictions. To do so, we compared the same set of classification metrics for the responses of two annotators in Chen-2024 against the LLM prediction. This showed that some relations are inherently difficult for human annotators and LLMs (e.g., S-targetstate, S-function). Other relations, like subordinates, have high agreement among annotators but low agreement in LLMs (see Figure 4) To illustrate, a pair like *sister – daughter* is coded as a subordinate relation. At least two factors could potentially explain these findings. First, in most cases, synonyms and antonyms

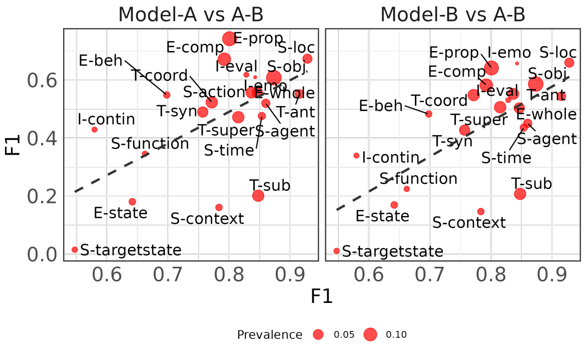


Figure 4: Comparing Coder A and B F1-scores vs Model predictions. Node size indicates prevalence. Observations under the regression line show relations that have higher F1 for human coders compared to LLM classifications against Coder A (left panel) or Coder B (right panel).

are also coordinates. As such, this suggests a shortcoming of the WB ontology, which could be resolved by adding a third level to the hierarchy where synonyms/antonyms are nested under coordinates. Second, GPT-4 might struggle with directional relations such as superordinates and subordinates, which is supported by the patterns in Figure 2 and Figure 3, showing difficulties distinguishing between synonyms, superordinates and subordinates. To investigate this possibility in more detail, we prompted GPT-4 with explicit propositions such as *daughter is a kind/type/instance of sister*, but this did not dramatically improve results.

While responses in the WAT are ambiguous without further insight from the participant who generated them, it is possible that in cases with ambiguity, associations are prone to several biases that promote certain interpretations over others. Specifically for concrete words, our results contrasting the Vivas-2022 features with association responses that removed relation indicators suggest that for concrete concepts, association-like features can be generated without much loss of information.[3] However, performance also depends on concreteness. The high performance for the Vivas-2022 association dataset might reflect the fact that most of the words were very concrete. However, consistent with previous findings by Liu et al. (2022), the Bolognesi-2017 metrics, which cover both concrete and abstract concepts, were somewhat lower than the primarily concrete data from Vivas-2022.

[3]One caveat is that the PLT is a more restricted form of the WAT because only a subset of semantic relations are highlighted in the participant's instructions (often accompanied by examples), whereas word associations are free.

4.3. Limitations

The use of a closed-sourced model has several inherent limitations. While these have been discussed at length elsewhere (e.g., Frank, 2023), it should be noted that some limitations are practical in nature. One of them is that different prompting regimes cannot be controlled experimentally as the cost to do so would become prohibitively large. In all our analyses, the model was asked to generate responses for 100 item pairs simultaneously, reflecting such constraints.

Second, there are also limitations to the WB ontology. On the one hand, some of the distinctions used in the original work were specific to research questions related to groundedness (e.g., in contrasting internal and external perceptual features, as these were implied in mental simulations) (Wu and Barsalou, 2009). The ontology also needs to be further adapted to work for word associations. While this does not present major difficulties, some details do not translate well (e.g., "Contingencies and complex cognitive operations"). Furthermore, distinctions between entity and situations properties, such as *function* (currently encoded as a Situational property) or *behavior* (currently encoded as an entity property), tend only to be distinguished in terms of how typical they are for an entity or a situation. As a consequence, some of the entity vs situation properties might be conflated with whether they apply in most situations or specific ones.

4.4. Future directions

The current work primarily focused on the WB ontology. Still, other ontologies have taken inspiration from modal-specific neuroscientific models to distinguish different ways in which words could be related (Garrard et al., 2001; Montefinese et al., 2013; Vinson and Vigliocco, 2008). It would be interesting to see how SOTA LLMs would account for these, especially since this would require access to accurate perceptual information (but see Marjieh et al., 2023, for a convincing demonstration of GPT-4 in this area).

An alternative approach could *infer* task-specific relation ontologies from word associations themselves. Liu et al. (2022) collected free-text explanations with word associations and then clustered explanations into data-driven relation types without supervision. LLMs may be prompted with a less constrained framework to allow for the generation of a label inventory from scratch.

While the current work focuses on labelling a single relation, the ontologies allow for multiple relation labels for a specific cue-response pair. A more refined procedure would consider the possibility that multiple labels might apply but vary in degree or prototypicality (Jurgens et al., 2012; Liu et al., 2022). Here one possibility would be to derive classification probabilities from a fine-tuned LLM in combination with either a sparsity constraint or a rule-based approach to ensure the number of relations that can be inferred remains small. Furthermore, much more work is also needed to determine the best way to prompt the model, including which definitions to give and what examples to provide (see Jurgens et al., 2012, for an interesting analogy-based approach). Furthermore, it is likely that different types of LLMs benefit from different prompt types, and further gains could be achieved by, for example, implementing a voting mechanism across multiple LLMs.

More broadly, many questions remain about determining what semantic relations to derive in the first place. While an answer to this depends on the intended use of these relations, LLMs could assist us in iteratively refining existing ontologies by merging or splitting distinctions or refining definitions of relations. This could go in tandem with a data-driven use of LLMs to freely group different types of cue-response pairs or label the relations might prove useful (e.g., Liu et al., 2022).

5. Conclusion

Recent Large Language Models hold considerable promise in annotating semantic relations from human elicitation tasks such as word associations. The current results suggest that broad distinctions are adequately captured by GPT-4, which is considered state-of-the-art at the moment of writing. GPT-4 requires very limited requirements editing of responses, which is important to scale the approach. However, there is sufficient room for improvement, especially for more fine-grained distinctions, such as different types of taxonomic relations. While the recovery of latent semantic relations in word association data will always be subject to some degree of ambiguity, the current results also suggest several ways in which existing coding schemes can be improved to facilitate the annotation process, which ultimately would benefit the automatic labelling of these relations as well.

6. Acknowledgements

We want to thank the Complex Human Data Hub at the University of Melbourne for funding support, Chen Wen, Meredith McKague, and Enzo Capistrano for thoughtful discussions, and the reviewers for their valuable suggestions. LF is supported by the Australian Research Council Discovery Early Career Research Award (Grant No. DE230100761).

7. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Marianna Bolognesi, Roosmaryn Pilgram, and Romy van den Heerik. 2017. Reliability in content analysis: The case of semantic feature norms classification. *Behavior Research Methods*, 49:1984–2001.

Andrew Cattle and Xiaojuan Ma. 2017. Predicting word association strengths. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1283–1288.

Wen Chen, Chunhua Liu, Meredith McKague, and Simon De Deyne. 2024. Semantic alignment in Chinese and English: a concept ontology-based approach.

Simon De Deyne, Danielle J Navarro, Andrew 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.

James Deese. 1965. *The structure of associations in language and thought*. Johns Hopkins Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Tess Fitzpatrick and Peter Thwaites. 2020. Word association research and the l2 lexicon. *Language Teaching*, 53(3):237–274.

Michael C Frank. 2023. Openly accessible llms can help us to understand human cognition. *Nature Human Behaviour*, 7(11):1825–1827.

Peter Garrard, Matthew A Lambon Ralph, John R Hodges, and Karalyn Patterson. 2001. Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive neuropsychology*, 18(2):125–174.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(3).

Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. 2007. Topics in semantic representation. *Psychological review*, 114(2):211.

Simon Jerome Han, Keith J Ransom, Andrew Perfors, and Charles Kemp. 2024. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364.

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.

Chunhua Liu, Trevor Cohn, Simon De Deyne, and Lea Frermann. 2022. Wax: A new dataset for word association explanations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 106–120.

Chunhua Liu, Trevor Cohn, and Lea Frermann. 2021. Commonsense knowledge in word associations and ConceptNet. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 481–495.

Hongjing Lu, Nicholas Ichien, and Keith J Holyoak. 2022. Probabilistic analogical mapping with semantic relation networks. *Psychological review*, 5:1078–1103.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. 2023. What language reveals about perception: Distilling psychophysical knowledge from large language models. *arXiv preprint arXiv:2302.01308*.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2013. Semantic memory: A feature-based analysis and new norms for italian. *Behavior research methods*, 45:440–461.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Massimo Stella, Nicole M Beckage, and Markus Brede. 2017. Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific reports*, 7(1):46730.

Sean Trott. 2023. Can large language models help augment english psycholinguistic datasets?

David P Vinson and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190.

Leticia Vivas, Maria Montefinese, Marianna Bolognesi, and Jorge Vivas. 2020. Core features: measures and characterization for different languages. *Cognitive processing*, 21(4):651–667.

Leticia Vivas, M Yerro, Sofía Romanelli, A García Coni, Ana Comesaña, F Lizarralde, I Passoni, and J Vivas. 2021. New spanish semantic feature production norms for older adults. *Behavior Research Methods*, pages 1–17.

Lingling Wu and Lawrence W Barsalou. 2009. Grounding concepts in perceptual simulation: Evidence from property generation. *Acta Psychologica*.

What's in a Name? Electrophysiological Differences in Processing Proper Nouns in Mandarin Chinese

Bernard Jap[1], Yu-Yin Hsu[2], Lavinia Salicchi[2], and Yuxi Li[2]
University of Macau[1], The Hong Kong Polytechnic University[2]
bernardjap@um.edu.mo, yyhsu@polyu.edu.hk, {lavinia.salicchi, yu-xi.li}@connect.polyu.hk

Abstract

The current study examines how proper names and common nouns in Chinese are cognitively processed during sentence comprehension. EEG data was recorded when participants were presented with neutral contexts followed by either a proper name or a common noun. Proper names in Chinese often consist of characters that can function independently as words or be combined with other characters to form words, potentially benefiting from the semantic features carried by each character. Using cluster-based permutation tests, we found a larger N400 for common nouns when compared to proper names. Our results suggest that the semantics of characters do play a role in facilitating the processing of proper names. This is consistent with previous behavioral findings on noun processing in Chinese, indicating that common nouns require more cognitive resources to process than proper names. Moreover, our results suggest that proper names are processed differently between alphabetic languages and Chinese language.

Keywords: proper names, common nouns, ERP, Mandarin Chinese

1. Introduction

A longstanding philosophical and linguistic debate concerns the definition and distinction between proper names (PNs) and common nouns (CNs). CNs are widely accepted as words that denote *classes* of real-world concrete and abstract entities, while PNs specify a particular *individual* entity within a class (Yasuda et al., 2000). Their difference is clear in clinical settings and daily life. For example, aphasic patients may struggle to recall either CNs or PNs (Warrington and McCarthy, 1987). Systematic difficulties in recalling PNs has also been related to early stages of Alzheimer's disease (Mueller et al., 2020; Semenza et al., 2003). These findings suggest that PNs pose more processing challenges (Brown, 1991). Adorni et al. (2014) propose that one key processing difference between the two categories is the relationship between a word and its reference. CNs are associated with a wider range of real-world entities and contain richer semantic features, whereas PNs have a direct connection to individual entities and are thus mostly associated with episodic memory.

Nonetheless, most of this research on PNs and CNs has been conducted in languages that use the Latin alphabet, such as English and Italian. These languages follow precise orthographic rules to distinguish words in the two categories, i.e., capitalizing the first letter for PNs (Sulpizio and Job, 2018). Considering this, we think Chinese language, a typologically distinct language, provides a useful context for further understanding the differences in processing PN and CNs. Chinese writing is mostly logographic, offering no orthographic or typographical cues to distinguish nouns' subcategories. Moreover, Chinese characters do not exclusively represent sounds; they also convey the semantics of the concepts they symbolize. For example, the two characters in the name 老周(Lao Zhou), 老and 周, can each be used independently as a standalone word, e.g., 'old' and 'week', respectively, while they may have other meanings in specific contexts. Even in PNs where the characters are not independent words, the fact that these characters are used in other words and may preserve some senses from those words potentially makes the processing of Chinese PNs different from that of PNs in alphabetic languages, in which individual letters of a PN (like *John*) do not have semantics.

Nonetheless, previous studies on processing Chinese PN and CNs, adopting either ERP (Wang et al., 2016) or behavioral method (Yen and Müller, 2003), do not reach a consensus. The question remains as to whether a language that does not impose orthographic constraints in distinguishing PNs from CNs, like Chinese, exhibits different processing patterns during the language comprehension of CNs and PNs, in which the latter type also contains semantically meaningful characters. Therefore, to address this question, we conducted an ERP experiment to investigate the processing of PNs and CNs in Mandarin during language comprehension.

2. Related Work

Lu and Bai (2023) examined whether CNs and PNs are processed differently in the left and right hemispheres by Chinese speakers. Their results sug-

The Workshop on Cognitive Aspects of the Lexicon, pages 79–85
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

gest a lateralization of CN processing, while PNs did not show the same hemispheric advantage. However, these findings are partially disproved by Desai et al. (2023), which showed that in fMRI both PNs and CNs activated a wide network, across both hemispheres, with several overlapped active areas. These differences concerned the level of activation of these areas, with PNs leading to greater involvement of the right hemisphere.

Further proof of the processing differences between PNs and CNs can be found in EEG studies. Dehaene (1995) tested the neural correlates of five sub-categories of nouns and found a stronger N400 in temporal regions associated with PNs. Adorni et al. (2014) recorded in a lexical decision task the same late negativity with a P300 linked, again, to PNs only. Early correlates were recorded as well: in Müller and Kutas (1996) and Proverbio et al. (2001) a stronger P200 and N100 in the left anterior temporal and left fronto-central cerebral areas were found in association with PNs. Proverbio et al. (2009) focused on the evaluation of pairs of words being either name/surname or compounds of CNs, both existent and nonexistent while reaction time (RT) and brain activities were recorded. In contrast with previous studies, it revealed longer RTs and a stronger N400 in association with CNs.

Most previous research focused on PNs and CNs has been conducted in languages that use the Latin alphabet and require capital first letter for PNs. Sulpizio and Job (2018) studied the influence of orthographic variations on noun processing and found that N100 and P200 are associated with early processing of the form-category typicality. This indicates that studies on alphabet-based languages are heavily influenced by orthographic rules.

Wang et al. (2016) presented to Chinese participants PNs and CNs that were (in-)congruent with a given context. The ERP analysis showed that the N400 elicited by incoherent sentences was stronger in front of PNs, especially in the left hemisphere. Incongruent CNs, however, led to a stronger P600, suggesting a later, and more challenging, processing. Finally, Yen and Müller (2003), a behavioral study on Chinese nouns, found that CNs were harder to process than PNs, leading to longer RTs.

Generally speaking, there is no consensus on the processing of CNs and PNs, and most of them studied nouns independent of contexts. Previous studies do not agree on the brain networks involved in processing the two noun categories, whether or not lateralization takes place, what the neural correlates associated with PNs and CNs are, and the timing of their process. However, PNs seem to be more challenging in alphabetic languages.

In this study, we examined the processing patterns of PN and CN in Mandarin sentence comprehension, to simulate real-world language processing as both noun types are often encountered in reading or listening. We hypothesize that if semantic access to PNs through individual characters indeed facilitates their reading, processing PNs should require similar or even less cognitive effort compared to CNs, as suggested by Yen and Müller (2003). Specifically, we examine early potentials (N100, P200), and the N400. Were there no observable ERPs *OR* observable N100, P200, or N400 for the CN stimuli, it can be interpreted as evidence for semantic facilitation in processing Chinese PNs. Conversely, if observable N100, P200, or N400 for the PN stimuli occur, it would suggest that Chinese PNs are processed similarly to PNs in Indo-European languages, where the sub-components typically do not correspond to a lexical entry.

Table 1: An example set of the target nouns. *Note: PN means PN, and AN means animate nouns.*

	Context sentence	Target
PN	在学校组织的郊游途中， 'In a trip organized by the school,'	小婷.... 'Xiaoting'
AN	在学校组织的郊游途中， 'In a trip organized by the school,'	妹妹.... 'sister'

3. Method

3.1. Stimuli

As each experimental item started with a neutral sentence context, we divided nouns into 24 PNs and animate 24 CNs, with animacy effect in control. Before the experiment, a naturalness judgment task of the experimental items was conducted on a five-point Likert scale by 30 native speakers of Mandarin who did not participate in the experiment. All items used in this study were rated as 3 points or above in the judgment task. While we did not set a specific parameter for the selection of PNs and CNs, all PNs and CNs are two-character Chinese words/names, and their linguistic characteristics are delineated in subsection 3.3. Each set of experimental sentences (Table 1) involves two types of target nouns serving as the subject of a sentence and was introduced by the same context sentence in each set. Sentences were pseudorandomized and organized into two sets; each set had 48 trials and 72 fillers. The materials were presented in simplified Chinese characters, and word-by-word, each for 600ms, with a 500ms blank screen between each word, except for the context sentence displayed as a unit for 2000ms. Digital triggers were manually inserted at the relevant time point in every sentence, which is at the onset of the noun.

3.2. EEG data

3.2.1. Participants and Procedure

47 adult native speakers (mean age 22.21 $\pm$ 2.35, 26 female) of Mandarin Chinese from the Northern provinces of China participated in our study. Data of 9 participants were excluded due to low trial counts that remain after artifact rejection; thus, 38 participants' data were used for analyses.

In each experiment, participants were seated in front of a monitor presented with the sentences using E-Prime. The monitor showed written instructions that were explained orally by the experimenter. Participants were instructed to minimize head movements and keep their eyes open during the experiment, but blinking was allowed. During the experiment, a fixation cross was shown between trials and sets. Each trial began with a 500ms blank screen, followed by the phrase "准备好了吗？" (ready?)" shown on the screen until participants pressed any key. To prevent fatigue, there were breaks after every block (10 blocks in total), allowing the participants to read at their own pace. After every 3-8 trials (randomized), a comprehension prompt was given to ensure that the participants remained focused and to provide a measure of comprehension performance. Each session lasted about one hour, including cap and electrode preparation. Participants received US$25.

3.2.2. Measurements and Preprocessing

The participants' EEGs were recorded using a 64-channel and then preprocessed using EEGLAB (Delorme and Makeig, 2004). FieldTrip (Oostenveld et al., 2011) was used for statistical analysis. The EEG data was re-referenced to the two mastoid electrodes, and bad channels were interpolated. The remaining data was filtered using a 0.1Hz high-pass filter. ERPs were calculated for each participant, electrode, and condition in an interval from 200ms before onset to 1000ms after onset for each time-locked trigger. These epochs were then demeaned per channel and subjected to independent component analysis. Components associated with blinks, saccades, and muscle artifacts were removed. After this step, baseline correction was applied to the data using a 200 ms pre-stimulus onset baseline. Then, a threshold rejection function was used to detect and reject artifacts. Finally, data was filtered using a 40.0Hz low-pass filter.

3.2.3. Statistical analysis

For the statistical analysis, we used cluster-based permutation tests (Maris and Oostenveld, 2007) on all the scalp electrodes and the specific time window within the epoch. The tests compared the PN and CN ERPs at each channel and each sample, identifying clusters of spatiotemporally adjacent data points where the difference between the two conditions exceeds a threshold of $p < .1$ in a t-test for that time window. Given the previous ERP results, we conducted two cluster tests to capture potential differences between PNs and CNs: (1) a two-tailed test on 0-300ms to measure early potentials that have been reported, specifically N100 and P200. (2) a two-tailed test on 300-500ms to measure the N400, which has been the most common ERP observed in the study of PN processing. In (2) we ran a two-tailed test despite testing for an ERP of negative polarity because the direction of the effect is unclear, as we reviewed in Section 2.

3.3. Extraction of linguistic features

For each target noun and its initial and final characters, we extracted four types of linguistic features: frequency, stroke counts, word status of the characters, and finally the orthographic neighbor density of the noun. The purpose of extracting these features is to better interpret the ERP results in light of the differences in linguistic features between CN and PN: were we to find effects resembling, for example, a frequency effect in a direction that fits the word frequency profile of the stimuli, we would be equipped to avoid misinterpreting specific ERPs as waveforms unique to processing that is associated with common or proper nouns.

We calculated **frequencies** based on the corpus 'Chinese Web 2017 (zhTenTen17) Simplified' in Sketch Engine (Kilgarriff et al., 2014). We then normalized the frequencies on the basis of 10,000 words. We retrieved the **stroke number** of characters from hanziDB [1].The **word status** of characters was determined by two of the authors of this study who are native speakers of Mandarin, and specialized in Chinese linguistics. The criteria were the characters' meaningfulness and independence. Following Xiong et al. (2021), we calculated the **orthographic neighbors**, and its **density** (as the ratio of orthographic neighbors to the number of total word types). Table 2 (Appendix A) shows the means for these features.

Word frequency and visual complexity have been shown to impact early ERPs. A meta-study of 1100 English words and pseudowords by Dufau et al. (2015) observed frequency effects starting from 100ms, while visual complexity affects begin as early as 30-50ms, with another time window at 100-150ms also showing such effect. In the case of Chinese characters, visual complexity may be assessed by stroke count of a word or a character. Characters with more strokes elicited larger P200 and smaller N200, and similar levels of N400

[1] `http://hanzidb.org/`

(Yang et al., 2016). Similarly, words with many orthographic neighbors generate larger N400 than words with fewer ones (Müller et al., 2010).

In our stimuli, the word frequency for CNs is much higher than for PNs, which is expected when extracting generic names from corpora. However, PNs have higher individual character frequency. The stroke count is comparable, with the first character of PNs having a lower stroke count. The orthographic neighborhood shows that PN stimuli have a higher orthographic neighborhood compared to the CN stimuli. There is strong evidence that factors like frequency are task-dependent (Fischer-Baum et al., 2014), and not all tasks will elicit ERPs that correspond to these factors. Therefore, it is possible that these effects will not be observed in our study, despite the items not being perfectly balanced for these factors.

4. Results

Figure 1 shows grand average waveforms at electrodes starting from the baseline (200ms before the stimuli onset) to 1000ms. The contrast between CNs and PNs generated a clear negative shift at around the time window of the N400, peaking at almost precisely 400ms. Additionally, a minor contrast seems to occur at around the 250ms mark with the PN condition having a higher peak amplitude.

Specific channels were chosen to present the results when the channels corresponded to significant differences in the cluster test (refer to the raster plot in Appendix B). Figure 2 shows topographic maps for the waveforms of CN minus PN between 300ms and 500ms. The color bars on the right side of the topographic plots show the amplitude of the channel from -2µV (blue) to 2µV (yellow). The effect's topographical distribution intensifies between 350-450ms, with the negativity ranging from centroparietal to occipital channels. The ERP is centered in the midline channels, distributed along both sides of the hemisphere. The latency at around 300-500ms, the peak of the ERP at around 400ms, and the central to posterior distribution of the effect suggest that this is an N400. Although the topography seems to be distributed in a slightly more posterior area compared to the typical centroparietal N400, significant differences were found at the centroparietal channels and even some of the central electrodes (Figures 1-3).

Statistical analyses confirmed these observations. In the early potentials time window (0-300ms) for the N100 and P200, we find no effect, but a significant (p=.028) effect at the N400 time window (300-500ms) across 28 channels from central to occipital electrodes with the negative shift for CNs starting from 332ms to 484ms. We also found the clusters in the permutation test for the 300-500ms time window to correspond with the ERP and topographic plots (see Figure 3 in Appendix B). Our results show an N400 with a mainly central to posterior distribution for CNs when compared to PNs.

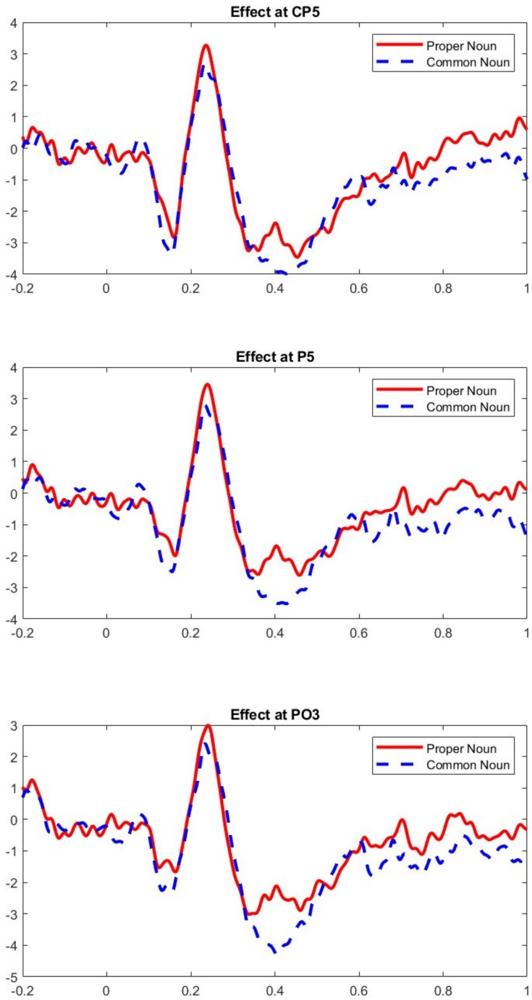


Figure 1: ERPs at three individual channels.

5. Discussion and Conclusion

The representational model proposed by Cohen (1990) postulates that the difficulty in retrieving PNs stems from processing rather than storage, as names are typically semantically neutral and offer little semantic clues for retrieval. However, this may not be fully apply to all languages. In Chinese, many names' characters can function as a standalone word. In our stimuli, the majority of the PNs (83 percent) contain such characters. Thus, Chinese readers are likely to use the semantic information of such characters to facilitate PN processing. However, if this is the case, we should expect *no* differences between PNs and CNs, since both types allow semantic facilitation in Chinese. This

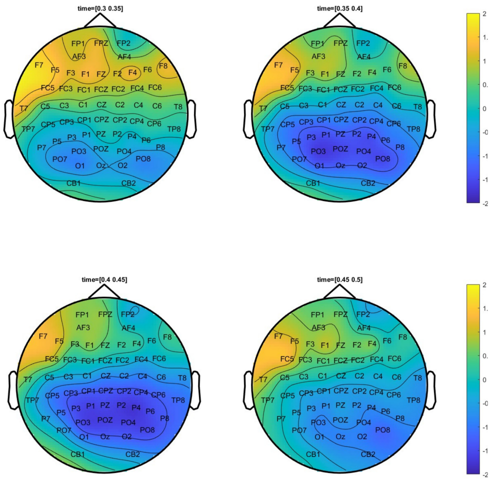


Figure 2: Topographic plot for CNs - PNs from 300ms to 500ms in 50ms intervals.

raises the question of why the differences still exist.

Our results differ from the ERP study by Wang et al. (2016) that reported larger N400s for incongruent PNs compared to CNs; yet, it is important to note that the differences in that study were elicited from a violation-based design, making comparisons between the two studies inappropriate. While limited evidence directly supports the idea that each of the various factors (e.g., frequency and imageability) facilitates the processing of either type of nouns, some studies suggest that most PNs' lack of semantic features is the primary factor to PNs' increased processing efforts (see Adorni et al. (2014) for a summary). Our finding is consistent with results in Yen and Müller (2003), showing that CNs are more difficult than PNs.There are several possibilities as to why PNs can be easier to process than CNs in Chinese: a) PNs are usually more imageable than CNs (Proverbio et al., 2009); b) PNs may evoke more emotional and sensory activations than CNs (Gorno-Tempini et al., 1998; Douville et al., 2005) with the left temporal cortex playing an important role in PN retrieval; c) the retrieval of PNs generates visual representation in brain areas involved in processing of visual images, even when not required by the tasks (Campanella et al., 2001), and in this case, visual imagery generation may assist in the processing of PNs.

Limitations

The main limitation is that there are linguistic characteristic differences between the two noun groups: compared to CNs, the PNs in the stimuli have a lower frequency and more orthographic neighbors, but the two groups have somewhat comparable stroke counts. In this case, according to previous research, we would expect to find early effects from 100ms (Dufau et al., 2015) which corresponds to low frequency, and a larger N400 amplitude (Müller et al., 2010) which corresponds to higher orthographic density for the PN. Nonetheless, in our data, we did not find ERPs that correspond with word frequency and neighborhood density. We did not find any early potentials, and instead, we observed an N400 for the CNs, which are words with higher frequency. As such, it is rather unlikely that the effects found in the analysis are driven by differences in the listed linguistic features, given that their corresponding ERP effects are not present.

Acknowledgment

This study was approved by the University's Ethics Review Board (HSEARS20211223003), and was supported by the Departmental General Research Fund (4-ZZRX) funded by the Hong Kong Polytechnic University. We would like to thank the anonymous reviewers for their feedback and suggestions. Special thanks to Deran Kong, Wenxi Fei, and Zhihong Chen for assisting in part of the experimental preparation.

Bibliographical References

Roberta Adorni, Mirella Manfredi, and Alice Mado Proverbio. 2014. Electro-cortical manifestations of common vs. proper name processing during reading. 135:1–8.

Alan S Brown. 1991. A review of the tip-of-the-tongue experience. *Psychological bulletin*, 109(2):204.

Salvatore Campanella, Frédéric Joassin, Bruno Rossion, Anne De Volder, Raymond Bruyer, and Marc Crommelinck. 2001. Association of the distinct visual representations of faces and names: A pet activation study. *NeuroImage*, 14(4):873–882.

Gillian Cohen. 1990. Why is it difficult to put names to faces? *British Journal of Psychology*, 81(3):287–297.

Stanislas Dehaene. 1995. Evidence for category-specific word processing in the normal human brain. *NeuroReport*, 6(2):2153–2157.

Arnaud Delorme and Scott Makeig. 2004. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21.

Rutvik H. Desai, Usha Tadimeti, and Nicholas Riccardi. 2023. Proper and common names in the semantic system. 228(1):239–254.

Kelli Douville, John L Woodard, Michael Seidenberg, Sarah K Miller, Catherine L Leveroni, Kristy A Nielson, Malgorzata Franczak, Piero Antuono, and Stephen M Rao. 2005. Medial temporal lobe activity for recognition of recent and remote famous names: an event-related fmri study. *Neuropsychologia*, 43(5):693–703.

Stéphane Dufau, Jonathan Grainger, Katherine J Midgley, and Phillip J Holcomb. 2015. A thousand words are worth a picture: Snapshots of printed-word processing in an event-related potential megastudy. *Psychological science*, 26(12):1887–1897.

Simon Fischer-Baum, Danielle S Dickson, and Kara D Federmeier. 2014. Frequency and regularity effects in reading are task dependent: evidence from erps. *Language, cognition and neuroscience*, 29(10):1342–1355.

Maria Luisa Gorno-Tempini, Cathy J Price, Oliver Josephs, Rik Vandenberghe, Stefano F Cappa, Narinder Kapur, Richard S Frackowiak, and ML Tempini. 1998. The neural systems sustaining face and proper-name processing. *Brain: a journal of neurology*, 121(11):2103–2118.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1:7–36.

Zijia Lu and Xuejun Bai. 2023. The processing differences between chinese proper nouns and common nouns in the left and right hemispheres of the brain. 13(3):424.

Eric Maris and Robert Oostenveld. 2007. Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190.

Kimberly D Mueller, Rebecca L Koscik, Lianlian Du, Davide Bruno, Erin M Jonaitis, Audra Z Koscik, Bradley T Christian, Tobey J Betthauser, Nathaniel A Chin, Bruce P Hermann, et al. 2020. Proper names from story recall are associated with beta-amyloid in cognitively unimpaired adults at risk for alzheimer's disease. *Cortex*, 131:137–150.

Horst M Müller and Marta Kutas. 1996. What's in a name? electrophysiological differences between spoken nouns, proper names and one's own name. *NeuroReport*, 8(1):221–225.

Oliver Müller, Jon Andoni Duñabeitia, and Manuel Carreiras. 2010. Orthographic and associative neighborhood density effects: What is shared, what is different? *Psychophysiology*, 47(3):455–466.

Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. 2011. Fieldtrip: open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011:1–9.

Alice Mado Proverbio, Stefania Lilli, Carlo Semenza, and Alberto Zani. 2001. ERP indexes of functional differences in brain activation during proper and common names retrieval. 39(8):815–827.

Alice Mado Proverbio, Serena Mariani, Alberto Zani, and Roberta Adorni. 2009. How are 'barack obama' and 'president elect' differentially stored in the brain? an ERP investigation on the processing of proper and common noun pairs. 4(9):e7126.

Carlo Semenza, Sara Mondini, F Borgo, M Pasini, and MT Sgaramella. 2003. Proper names in patients with early alzheimer's disease. *Neurocase*, 9(1):63–69.

Simone Sulpizio and Remo Job. 2018. Early and multiple-loci divergency of proper and common names: An event-related potential investigation. 119:107–117.

Lin Wang, Rinus G. Verdonschot, and Yufang Yang. 2016. The processing difference between person names and common nouns in sentence contexts: an ERP study. 80(1):94–108.

Elizabeth K Warrington and Rosaleen A McCarthy. 1987. Categories of knowledge: Further fractionations and an attempted integration. *Brain*, 110(5):1273–1296.

Jianping Xiong, Yujie Zhang, and Ping Ju. 2021. The effects of orthographic neighborhood size and the influence of individual differences in linguistic skills during the recognition of chinese words. *Frontiers in Psychology*, 12:727894.

Shasha Yang, Shunmei Zhang, and Quanhong Wang. 2016. P2 and behavioral effects of stroke count in chinese characters: Evidence for an analytic and attentional view. *Neuroscience letters*, 628:123–127.

Kiyoshi Yasuda, Bobbie Beckmann, and Tetsuo Nakamura. 2000. Brain processing of proper names. 14(11):1067–1089.

Huei-Ling Yen and Horst M. Müller. 2003. Processing of proper names in mandarin chinese. In Franz Schmalhofer, Richard Young, and Graham Katz, editors, *Proceedings of EuroCogSci 03*, 1 edition, pages 450–450. Routledge.

Appendix A. Summary of Linguistic Features

	Freq	C1F	C2F	C1S	C2S	C1R	C2R	ON*
PN	0.01	5.11	0.65	5.29	7.96	0.58	0.46	128.62
AN	0.25	0.59	0.26	7.16	7.25	0.92	0.83	13.19

Table 2: Means of the targets' frequency (Freq), characters' frequency (C F), stroke count (S), word ratio (R), and orthographic neighbor (ON) (* in thousands)

Appendix B. Raster Plot

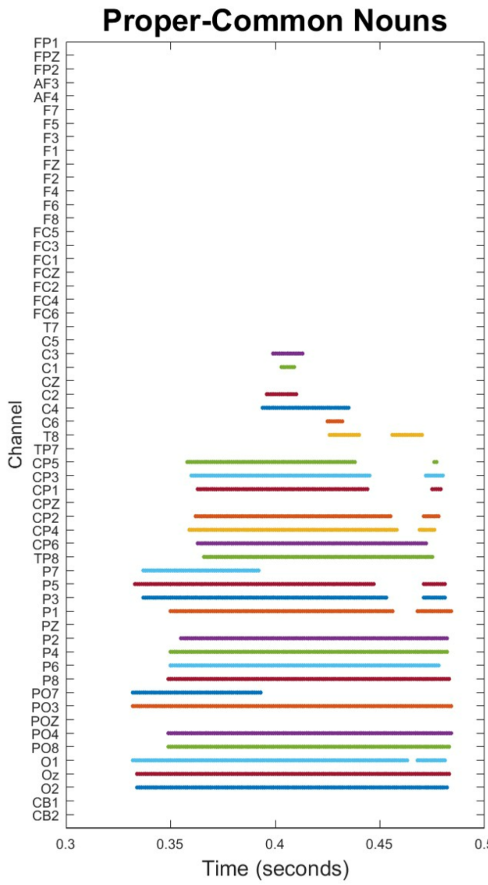


Figure 3: Raster plot showing data point included in the cluster tested with the permutation test for the 300-500ms time window

Cross-Linguistic Processing of Non-Compositional Expressions in Slavic Languages

Iuliia Zaitova, Irina Stenger, Muhammad Umer Butt and Tania Avgustinova
Department of Language Science and Technology, Saarland University, Germany
izaitova@lsv.uni-saarland.de, ira.stenger@mx.uni-saarland.de,
mubu00001@stud.uni-saarland.de, avgustinova@coli.uni-saarland.de

Abstract

This study focuses on evaluating and predicting the intelligibility of non-compositional expressions within the context of five closely related Slavic languages: Belarusian, Bulgarian, Czech, Polish, and Ukrainian, as perceived by native speakers of Russian. Our investigation employs a web-based experiment where native Russian respondents take part in free-response and multiple-choice translation tasks. Based on the previous studies in mutual intelligibility and non-compositionality, we propose two predictive factors for reading comprehension of unknown but closely related languages: 1) linguistic distances, which include orthographic and phonological distances; 2) surprisal scores obtained from monolingual Language Models (LMs). Our primary objective is to explore the relationship of these two factors with the intelligibility scores and response times of our web-based experiment. Our findings reveal that, while intelligibility scores from the experimental tasks exhibit a stronger correlation with phonological distances, LM surprisal scores appear to be better predictors of the time participants invest in completing the translation tasks.

Keywords: non-compositionality, closely related languages, language models, surprisal, linguistic distances

1. Introduction

The principle of compositionality in linguistics states that the meaning of a complex expression is determined by the meanings of its constituent parts (Partee, 2008). However, non-compositional expressions deviate from this principle. Non-compositional expressions are linguistic constructs where the overall meaning cannot be straightforwardly inferred from the meanings of their individual components (Baldwin and Kim, 2010). The meaning of non-compositional expressions often relies on cultural, contextual, or conventional associations, making them an aspect of language that requires specialized analysis beyond the scope of compositional interpretation (Jackendoff, 2002). Examples of non-compositional expressions include idioms (*e.g., English: "to kick the bucket" meaning: to die*), metaphors (*Czech: "Život je cesta", meaning "Life is a journey"*), and microsyntactic units (*Bulgarian:* "не веднъж" *transliterated as "ne vedn"ž"*[1]*, meaning "not once"; Russian:* "в конце" *transliterated as "v konce", meaning "at the end of"*).

While the mechanisms underlying the comprehension and processing of non-compositional expressions within a single language have been investigated extensively (Cacciari and Tabossi, 1988; Conklin and Schmitt, 2008; Titone et al., 2019), the dynamics of cognitive processing of written non-compositional expressions across languages – especially within closely related language groups like Slavic languages, remain a subject for exploration.

In light of this, the current study addresses the following research questions:

- **RQ1:** How well can native Russian speakers spontaneously understand non-compositional expressions from unfamiliar Slavic languages, namely Belarusian (BE), Bulgarian (BG), Czech (CS), Polish (PL), and Ukrainian (UK) in written context?
- **RQ2:** To what extent do algorithmic factors, namely surprisal from Language Models and linguistic distances, predict the cross-lingual intelligibility of non-compositional expressions?

The paper is structured as follows: we start by providing information on previous research in non-compositionality and language intercomprehension (Section 2) and stating our hypotheses (Section 3); then we describe our web-based experiment (Section 4) and algorithmic predictors (Section 5) to finally present (Section 6) and discuss the results in Section 7. The code for this paper is available at the following link: `https://github.com/IuliiaZaitova/non-compositional-expressions-slavic`.

2. Related Work

Spontaneous comprehension of unknown but related languages is detectable by means of differently designed experiments, e.g., cloze tests, multiple-choice questions, or translation tasks. For example, testing the Cyrillic script intelligibility by Russian native speakers in a context-free

[1] Here and further, we used ISO 9:1995 transliteration from Cyrillic.

The Workshop on Cognitive Aspects of the Lexicon, pages 86–97
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

word translation task, Stenger, 2019 reveals that Ukrainian and Belarusian are more understandable by the participants than Bulgarian, Macedonian and Serbian. The observed human performance in contextualized cross-lingual cognate recognition, as reported by Stenger and Avgustinova, 2021, also validates the intuition that Russian readers spontaneously understand stimuli in Ukrainian and Belarusian better than in Bulgarian.

When it comes to factors explaining the intercomprehension of related languages, researchers generally assume that the more similarities two languages share, the higher their degree of mutual intelligibility is (Gooskens and Swarte, 2017). As shown by Stenger and Avgustinova, 2021 linguistic distances are highly significant for correct in-context recognition of cognates from closely related languages. When looking at the intelligibility of Polish words to Czech readers, Jágrová et al., 2021 also confirms the role of linguistic similarity in predicting cross-lingual comprehension and finds that context-aware Language Models (LMs) perform better than 3-gram Language Models when predicting intercomprehension.

The exploration of different kinds of non-compositional expressions is fortified by a body of research consistently showing that these linguistic units exhibit increased processing facilitation (Cacciari and Tabossi, 1988; Conklin and Schmitt, 2008; Vespignani et al., 2009; Siyanova-Chanturia et al., 2011; Titone et al., 2019).

A relevant work by Kudera et al., 2023 investigates the auditory comprehension of idiomatic phrases, which is also a type of non-compositional expressions, in two closely related Slavic languages, Polish and Russian. The study builds on information-theoretic measures of word adaptation surprisal, coupled with syntactic distances between non-compositional expressions, to predict lay translators' preferences. Kudera et al.'s work serves as a foundational reference for our work; however, our approach diverges in several aspects: 1) we employ a reading comprehension scenario; 2) we test the comprehension in context; 3) we use five different target languages and compare the comprehension of non-compositional expressions across them.

A noteworthy study of the correlation between non-compositional expression intelligibility and LM performance is presented by Rambelli et al., 2023. Their work particularly focuses on idiomatic and high-frequency compositional expressions. The study indicates that humans process idioms with non-compositional meaning and high-frequency compositional phrases much faster than low-frequency compositional phrases. In parallel, LMs assign to idioms significantly lower surprisal values. In the context of our work, their findings underscore the potential of LM surprisal as a robust metric for predicting the processing of non-compositional expressions.

3. Hypotheses

RQ1: Our intention is to critically examine the alignment of our intelligibility tests with genealogic taxonomies established by comparative linguistics (Sussex and Cubberley, 2006), similarly to what is demonstrated in Charlotte Gooskens and Voigt, 2018. We hypothesize that native Russian speakers exhibit a higher comprehension level when exposed to non-compositional expressions in languages of the same East Slavic group (Belarusian and Ukrainian), and a lower comprehension level for languages in different groups (West Slavic and South Slavic). Moreover, we anticipate longer response times for languages more distant from Russian.

RQ2: Drawing upon previous studies in mutual intelligibility and non-compositionality, mainly Stenger and Avgustinova, 2021, Jágrová et al., 2021, Kudera et al., 2023, and Rambelli et al., 2023, we propose a dual-factor framework for predicting percentage of correct responses (intelligibility scores) and response times within our experimental context.

Factor 1: Linguistic Distances – we anticipate that more distant linguistic units will be more challenging for participants to recognize. Taking into account both orthographic and phonological distances, we predict a negative correlation between both types of linguistic distances and intelligibility scores.

Factor 2: Surprisal Scores from Language Models (LMs) – additionally, we incorporate surprisal scores from monolingual LMs trained on Russian. We analyze LM surprisal for 1) non-compositional Russian expressions in Russian context; 2) literal Russian expressions in Russian context; 3) non-compositional foreign expressions in foreign language context. We hypothesize a positive correlation between surprisal scores and user task completion time, with lower surprisal indicating processing facilitation. Additionally, we expect that surprisal scores of 1) and 2) correlate with results of multiple-choice question task since the low surprisal of the option in a particular context, which might be partially intelligible to the reader, can trigger the choice of that option (either literal or non-compositional). We also predict that 3) correlates with the outcomes of both tasks.

4. Human Translation of Unfamiliar Non-Compositional Expressions

In order to measure the intelligibility of non-compositional expressions we designed a two-task experiment that includes a free translation task and a multiple-choice task, each serving to probe different aspects of the participants' comprehension skills.

4.1. Stimuli

In our study, we utilize an existing dataset, initially crafted for the analysis of microsyntactic units, which are defined as non-compositional expressions with inherent syntactic idiomaticity. Such units include all the syntactic units that have very specific syntactic properties and do not fit into the standard syntax (Iomdin, 2015). The dataset consists of 227 Russian microsyntactic units, each accompanied by translational correlates and two parallel bilingual context sentences across six Slavic languages, as it is thoroughly described in Zaitova et al., 2023. The dataset was created using the Russian National Corpus (RNC) and its parallel sub-corpora as the primary linguistic resource (https://ruscorpora.ru). The microsyntactic dictionary provided by the RNC served as the pivot database. It includes various syntactic categories such as prepositions, adverbials, conjunctions, etc. The researchers selected the most frequent microsyntactic units for further analysis, totaling 227 units in Russian. Translational correlates were extracted from the RNC's parallel sub-corpora and the Czech National Corpus (Machálek, 2020), resulting in six parallel sets for each Slavic language under analysis. The dataset is open-sourced and available for use (https://huggingface.co/datasets/izaitova/slavic_fixed_expressions). While it was developed with a focus on microsyntactic units, in the current study we categorize these units as non-compositional expressions since, in line with the definition presented in Section 1, their meaning cannot be readily derived from their individual components. For each Slavic language in the dataset, we have selected a total of 60 expressions. The average sentence length in tokens per sentence is as follows: BE: 15.3, BG: 14.9, CS: 11.3, PL: 13.6, and UK: 14.8.

4.2. Word-by-Word Translation Options for Multiple-Choice Questions

A multiple-choice question format is employed in the experiment design as one the methods to assess participants' comprehension of presented non-compositional expressions. For each stimulus, participants are provided with a choice between two options: a correct translation and an literal translation counterpart, with the latter being crafted as a plausible yet inaccurate compositional translation of the respective expression, mirroring the stimulus in form. The goal is to challenge participants to discern between non-compositional (correct) and literal (incorrect) options. In the preparation of the assumed incorrect translations, we have utilized word-by-word translations provided by the online bilingual Glosbe Dictionary (https://glosbe.com). Additionally, for the identification of cognates, we use the etymological online dictionary of the Russian language by Max Vasmer (https://lexicography.online/etymology/vasmer/). The inclusion of literal translations as incorrect options aims at providing insights into participants' ability to move beyond surface-level comprehension and engage with the deeper (non-compositional) meanings of the investigated expressions.

4.3. Experimental Procedure

Cross-lingual intelligibility of non-compositional expressions to native Russian speakers has been assessed using a custom-built application available online at https://intercomprehension.coli.uni-saarland.de, as described by Stenger et al., 2020. The subjects received instructions in Russian about the tasks and procedures to follow. After familiarizing themselves with the task, participants registered on the website hosting our web application and completed a questionnaire about their background and language skills. During the experiment, participants saw five sets of 12 contextualized non-compositional expressions from one of the stimulus languages – Belarussian (BE), Bulgarian (BG), Czech (CS), Polish (PL), Ukrainian (UK). Each time, a set of 12 stimuli was randomly selected from all available sets per language, totalling 60 sentences per participant. Repetitions were avoided by ensuring that each stimuli set is presented to each participant only once. Stimuli sentences were presented one by one, and participants were first asked to type a free translation of the highlighted non-compositional expression (see Figure 1). Next, participants were presented with the multiple-choice question task (MCQ) task (see Figure 2) for the same stimulus. They were provided with two possible solutions for the translation of a foreign non-compositional expression into RU: (i) non-compositional translation; (ii) an alternative word-by-word translation, which is an inaccurate translation of the expression.

This combination of tasks was designed to be concatenated, with the addition of time limits to discourage lengthy reflection. While there may be some priming effect within the same stimulus, the difference between the two tasks (outlined in Section 4.5) does not appear to be primarily attributable to priming. Participants are presented more information in the multiple-choice options, leading to an expected increase in accuracy compared to the free translation task.

The time allocated for translating the highlighted non-compositional expression is based on a formula of 10 seconds per token plus an additional 3 seconds per sentence. For the second task, we add 3 more seconds plus 10 seconds per token in both translation options. Such timing is based on the experience with contextualized cognate guessing tasks and aligns with related studies, e.g., Stenger and Avgustinova, 2021. The timing of response for each stimulus starts when the question is shown to the user, and ends when the user proceeds to the next stimulus, either by providing a response or pressing the Skip button. For free translation task, we considered alternative semantically equivalent translations and typographical errors as correct responses. Accuracy in both tasks is defined as the percentage of correct responses out of total responses.

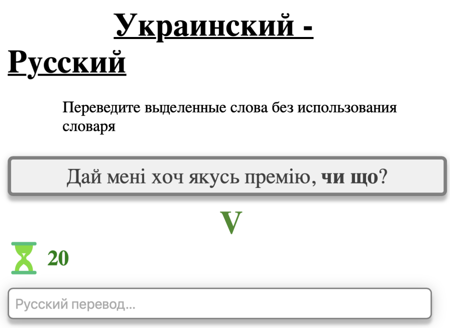


Figure 1: Experimental screen of the free translation task as seen by Russian respondents. The instruction reads: ‘Ukrainian - Russian. Translate the highlighted words without using a dictionary’. The Ukrainian sentence is: ’Give me at least some kind of bonus *or something*?’ The translation is to be written in the white box

4.4. Participants

In total, 135 native Russian participants took part in the study, aged between 20 and 78 years old (i.e. average age 35), comprising 92 females, 41 males, and 2 individuals who identified as another gender. The subjects were untrained in translation and were recruited for participation in the experiment through Prolific (`https://prolific.com`), an online platform specializing in participant recruitment for research purposes. To reveal the inherent intercomprehension, we excluded 12 participants because they had some knowledge of the stimulus language. Since the Prolific platform is in English, we expect that the speakers are familiar with the Latin script used by CS and PL languages. The number of subjects for each stimulus ranges from 17 to 55 with an average of 24 participants per stimulus. After each block, each participant may continue the experiment by completing the task for the remaining stimulus sentences offered in a random order.

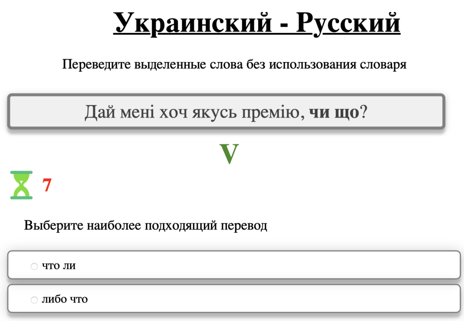


Figure 2: Experimental screen of the MCQ task as seen by Russian respondents. The instruction reads: ‘Ukrainian - Russian. Translate the highlighted words without using a dictionary’. The Ukrainian sentence is: ’Give me at least some kind of bonus *or something*?’ Below is the prompt: ’Choose the most suitable translation.’

4.5. Results

In Figure 3, the left plot illustrates the accuracy for both multiple-choice questions and free translation tasks, represented as the percentage of correct responses out of total responses. The right plot displays the response time for both tasks, organized by stimulus language. In both tasks, the highest accuracy is observed in translations from BE and UK. Since BE and UK belong to the same branch of Slavic languages as RU, such results are in line with the previous studies on Slavic language intercomprehension (Stenger and Avgustinova, 2021). Translations from BG also exhibited a relatively high accuracy. However, the accuracy dropped significantly for CS and PL. Generally, the participants’ performance is much lower in the free translation task, which is expected given that the task requires more open-ended and expressive language production.

As for time measurements, we can observe the opposite tendency: participants generally required more time when translating from BG, CS, and PL compared to BE and UK. This difference in time may reflect the additional effort and processing demands involved in comprehending and generating translations for languages that are less closely related or have greater linguistic differences.

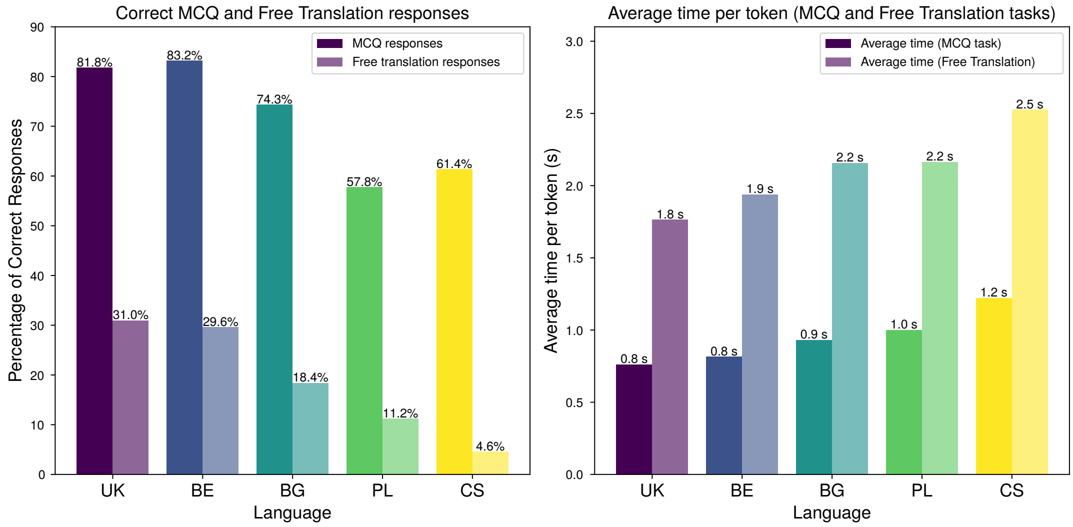


Figure 3: Left: Percentage of correct responses out of total responses (accuracy). Right: Average time per token in seconds. Both plotted by stimulus language

5. Predictors of Non-Compositional Expression Intelligibility

In this section, we describe the factors that we identified as predictors of our experiment metrics: linguistic distances and surprisal from LMs. In the section, we describe the two types of linguistic distances that we utilized and provide their comparative scores that further demonstrate their potential. We aim to investigate to what extent they can serve as a reliable proxy for cross-lingual intelligibility of non-compositional expressions in closely related languages.

5.1. Linguistic Distance

As outlined in Section 2, previous studies on intercomprehension provide strong support for using orthographic and phonological distances as a predictor of cross-lingual intelligibility (Vanhove and Berthele, 2015; Möller and Zeevaert, 2015; Gooskens and Swarte, 2017). However, measuring the distance between modern Slavic languages could be challenging due to the use of two writing scripts – Latin and Cyrillic. To accommodate for this, we employed two measures of phonological and orthographic distances that are adapted to deal with different scripts and were used before in Slavic intercomprehension studies specifically (Zaitova et al., 2023; Stenger et al., 2022; Mosbach et al., 2019).

5.1.1. Orthographic Distance

To measure the orthographic distance, we used **normalized Word Adaptation Surprisal (nWAS)**, which quantifies the degree of unexpectedness of a word form given a possibly related word form and set of transformation probabilities (Stenger et al., 2017). To use nWAS, orthographic character alignment costs are necessary. Based on these costs, words are aligned with the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). For our analysis, we adapted the code and the orthographic alignment costs previously computed for Slavic languages in Stenger et al., 2022. Here, identical characters have zero alignment cost, while characters differing only in diacritics (e.g., ⟨á⟩ and ⟨a⟩) were assigned a cost of 0.5. Unrelated vowel-vowel or consonant-consonant character pairs (e.g., ⟨a⟩ and ⟨i⟩, or ⟨k⟩ and ⟨v⟩) were assigned alignment costs of 1. Cyrillic hard and soft signs (⟨ь, ъ, ’⟩) were also assigned alignment costs of 1 to each other. For all other character pairs (e.g., consonant-vowel pairs), a cost of 4.5 was assigned. Cyrillic words were aligned by converting Cyrillic characters to ISO 9 Latin characters and then applying the alignment costs specified above.

5.1.2. Phonological Distance

Phonologically Weighted Levenshtein Distance (PWLD) is a measure of phonological similarity between different phonemic sequences or word forms (Fontan et al., 2016). The PWLD metric is an extension of the string-based Levenshtein distance that also takes into account the cost of each phoneme substitution based on phoneme features. These features are based on the PHOIBLE (Moran and McCloy, 2019) feature set. The substitution cost between phonemes is computed as the Hamming distance between their feature vector representations. We suppose that PWLD is more suitable for cross-lingual analysis than Levenshtein Distance since it is capable of catching less apparent phonological similarities. For example in the pair of Czech and Bulgarian cognates *ucho /u x o/* and ухо */u x ɔ/*, where phonemes */o/* and */ɔ/* are very similar to each other, PWLD would capture this similarity more effectively compared to Levenshtein Distance. We use the same adaption of the original PWLD proposed in Abdullah et al. (2021) to make it suitable for our analysis. To obtain the phonetic transcription of all stimuli and MCQ task options, we used CharsiuG2P, which is a transformer based tool for grapheme-to-phoneme conversion (Zhu et al., 2022).

It might seem counterintuitive that we consider phonological distance for written data. After all, native RU participants are not expected to know the correct pronunciation of the stimuli since they never learnt stimulus languages before. However, they can try to read stimulus aloud, i.e. try to understand unfamiliar languages using their inner speech (Alderson-Day and Fernyhough, 2015). Additionally, previous research has shown that a pronunciation-based distance is a better predictor of intelligibility than traditionally calculated orthographic distance (Jagrova, 2022).

5.1.3. Linguistic Distance Results

Table 1 presents the nWAS and PWLD scores, indicating the average distance from the correct non-compositional expression (NC) in RU to the source expression in the foreign language (L2). Additionally, it shows the distances from the inaccurate word-by-word translation (LIT) to foreign language (L2).

Language	Type	nWAS	PWLD
BG	LIT–L2	3.175	0.204
	NC–L2	**3.221**	**0.253**
BE	LIT–L2	3.236	0.213
	NC–L2	**3.249**	**0.220**
CS	LIT–L2	3.323	0.175
	NC–L2	**3.382**	**0.291**
PL	LIT–L2	3.332	0.208
	NC–L2	**3.389**	**0.298**
UK	LIT–L2	3.257	0.198
	NC–L2	**3.298**	**0.210**

Table 1: nWAS and PWLD scores

5.2. Surprisal from Language Models

Surprisal is a quantifiable measure of unpredictability, grounded in information theory (Crocker et al., 2016). Specifically, surprisal quantifies the negative log-likelihood of encountering a particular unit given its preceding context. The surprisal of a unit increases with decreasing probability, reflecting a higher degree of unexpectedness in a given linguistic context.

Surprisal from Language Models (LMs) serves as a proxy for the difficulty of cognitive processing of (foreign) non-compositional expressions in context. For sequential models like ruGPT3Large and ruGPT3Small, the probability of the expression given context is based solely on the left side, simulating reading from left to right. In contrast, for masked models like ruBERTa-large and ruBERTa-small, it considers both the left and right sides, simulating the utilization of the entire sentential context by the reader.

For example, let's take a sentence from the dataset that we used "_ _, что трассу полета можно менять только в интересах безопасности и защиты здоровья.." (transliteration: *"__, čto trassu polëta možno menât' tol'ko v interesah bezopasnosti i zaŝity zdorov'â..."*, translation: *"__ that the flight path can only be changed for safety and health protection."*) If the missing part is 'можно сказать' (transliteration: *"možno skazat'"*, translation: *"one cay say"*) and the surrounding context makes it highly expected, then the surprisal of the expression 'можно сказать' in this sentence would be low if one considers both left and right context (like masked models like ruBERTa-large and ruBERTa-small). If we consider only the nonexistent context left to the blank space, the model's surprisal would be higher as its uncertainty about the correct sequence of tokens increases.

The LM surprisal scores were obtained using the Python library minicons (Misra, 2022) for three scenarios:

- Surprisal of RU non-compositional expressions in RU context.
- Surprisal of RU literal expressions in RU context.
- Surprisal of foreign non-compositional expression in foreign context.

5.2.1. Language Models

We employ both large and small monolingual Russian LMs to compute surprisal values, using autoregressive models (ruGPT3Large and ruGPT3Small) and bidirectional models (ruBERTa-large and ruBERTa-small).

The LMs utilized in our experiments were developed by the SberDevices team[2] and are detailed as follows:

1. **ruBERTa-large (ruBL)** is an adaptation of the Roberta model (Liu et al., 2019), a masked model that was pre-trained on a substantial 250GB corpus of Russian text.

[2] https://sberdevices.ru

2. **ruGPT3Large (ruGPT3L)** is a large-scale sequential model based on the GPT-2 architecture (Radford et al., 2019).
3. **ruBERTa-small (ruBS)** is a smaller variant of the ruBERTa-large. While it maintains the robustness of its larger counterpart, ruBERTa-small offers a computationally less intensive alternative.
4. **ruGPT3Small (ruGPT3S)** is a scaled-down version of the ruGPT3Large model. The training process was designed to be more computationally efficient while pertaining the generation of linguistically rich and coherent text.

By employing models that utilize both sequential and masked prediction mechanisms, our experiments were designed to provide a full comparison and capture various aspects of language comprehension.

5.3. Surprisal Scores

Table 2 gives an overview of average surprisal scores of the RU non-compositional expressions in RU context (NC), literal RU expressions in RU context (LIT), and foreign non-compositional expression in foreign context (L2). In the last column, we can see the statistical significance of the difference between LIT and NC computed using the Wilcoxon signed-rank test. Additionally, Appendix A presents the boxplots for surprisal values from all stimuli. All the scores were derived from the models described above. We can see that the model ruBS does not detect any statistically significant difference between LIT and NC expressions. For that reason, we exclude this model from our predictors of intelligibility and response times.

	Model	NC	LIT	L2	LIT–NC
BG	ruGPT3S	3.916	7.597	9.333	***
	ruBS	14.549	14.540	14.918	NS
	ruGPT3L	3.646	7.662	9.540	***
	ruBL	1.013	7.496	10.511	***
BE	ruGPT3S	3.524	6.821	9.069	***
	ruBS	0.965	13.688	14.667	NS
	ruGPT3L	3.331	6.754	7.719	***
	ruBL	0.925	6.143	2.795	***
CS	ruGPT3S	3.758	7.258	14.388	***
	ruBS	14.660	15.305	23.778	NS
	ruGPT3L	3.695	7.329	13.743	***
	ruBL	1.140	7.004	10.783	***
PL	ruGPT3S	3.679	7.508	14.183	***
	ruBS	14.749	14.681	26.592	NS
	ruGPT3L	3.475	7.459	13.291	***
	ruBL	1.037	6.380	9.015	***
UK	ruGPT3S	3.570	7.270	8.599	***
	ruBS	14.411	14.424	14.886	NS
	ruGPT3L	3.394	7.284	7.412	***
	ruBL	0.850	6.510	1.946	***

*=p$< .05$, **=p$< .01$, ***=p$< .001$, NS=Not Significant

Table 2: LM surprisal + Wilcoxon signed-rank Test

6. Results

6.1. Correlation Results

We have computed the Pearson correlation of the percentage of correct responses and average response time in both tasks with orthographic and phonological distances, as well as with surprisal scores listed in Table 2. In Appendix B, you can find the tables with results for all Pearson correlations, along with corresponding p-values. For accuracy in free translation task, the strongest correlation is observed with phonological distance (PWLD) between Russian non-compositional expression and foreign non-compositional expression (BE: -0.405**, BG: -0.471***, CS: -0.361**, PL: -0.428***, UK: -0.606***).

For accuracy in MCQ task, there is also a strong correlation for PWLD for all languages except BE (BE: -0.229 NS, BG: -0.417**, CS: -0.283*, PL:, -0.385**, UK: -0.502***)[3]. Additionally, for BE and UK, we can observe a strong significant positive correlation of MCQ translation accuracy and PWLD between Russian literal expressions and foreign non-compositional expressions (BE: 0.429***, UK: 0.307*).

Figure 4 presents the correlation of free translation accuracy and PWLD between Russian non-compositional expressions and foreign non-compositional expressions for UK on the left, and the correlation of MCQ translation accuracy and PWLD between Russian literal expressions and foreign non-compositional expressions for BE on the right.

Average time measurements for both tasks have a stronger correlation with surprisal from LMs for foreign expression in foreign context in most languages, especially with that from the model ruBERTa-large (ruBL). For free translation time: BE: 0.443***, BG: 0.135 NS, CS: 0.547***, PL: 0.304*, UK: 0.217 NS. For MCQ time: BE: 0.457***, BG: 0.102 NS, CS: 0.452***, PL: 0.308*, UK: 0.215.

Figure 5 presents the correlation of free translation time and ruBL surprisal for foreign expression in foreign context for CS on the left, and the correlation of MCQ time and ruBL surprisal for foreign expression in foreign context for PL on the right.

It is worth noting that no statistically significant correlation was detected for the time measurements in the Ukrainian language.

[3] here and further: NS: Not Significant, *: $p < .05$, **: $p < .01$, ***: $p < .001$.

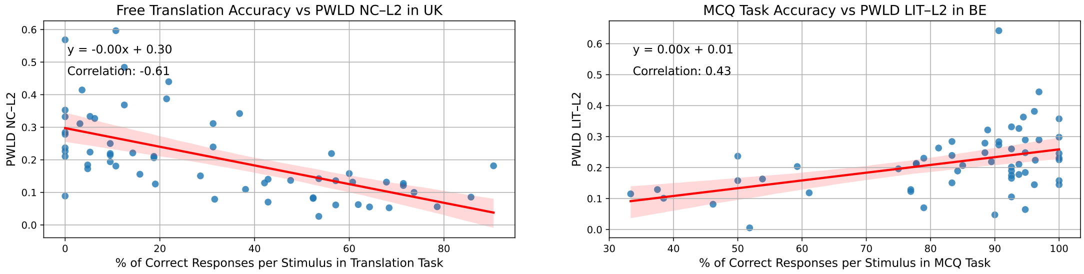


Figure 4: Relation of accuracy with phonological distances

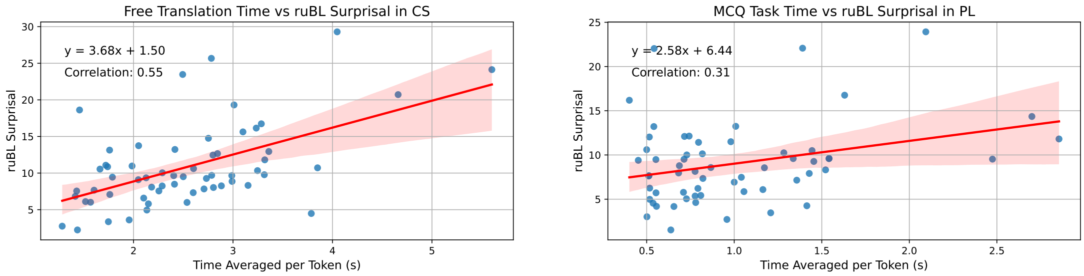


Figure 5: Relation of response time with LM surprisal

6.2. Multiple Regression Results

By adding all variables for surprisal and linguistic distances into a multiple linear regression model for predicting intelligibility across all stimuli from all source languages, we sought to identify the best-fitting models to predict intelligibility scores and average response times in our dataset. To achieve this, we performed a series of regression analyses using the Ordinary Least Squares (OLS) method.

We began by considering all potential predictor variables and identified the models that demonstrated the best fit for our data using stepwise regression. The summary of these results is presented in Table 3. Though overall regression scores are low when comparing the results for all language sets jointly, certain patterns could be observed across variables. For free translation (FT) task accuracy, the phonological distance between the correct RU translation and foreign stimulus (PWLD NC–L2), as well as surprisal scores for foreign stimulus in foreign context from models ruBL and ruGPT3L turned out to be the most significant predictors. For MCQ Task accuracy, phonological distances, namely PWLD NC–L2 and PWLD LIT–L2 (distances between the incorrect/literal translation option and foreign stimulus), again emerged as most impactful, followed by LIT surprisal by ruGPT3S and ruBL, and L2 surprisal by ruBL.

When considering response times, the best model for FT time includes ruBL L2, nWAS NC–L2, ruGPT3S L2, and ruBL LIT. Conversely, the MCQ Time model indicates that ruBL L2 and ruBL LIT are the most significant predictors, while nWAS orthographic distance does not have any significant impact. Phonological distances do not have any significant effect on both response time variables.

Dep. Variable	R^2	Adj. R^2	F	Variable	Coef
FT Accuracy	0.349	0.342	52.03	PWLD NC-L2	-72.4037
				ruBL L2	-0.8590
				ruGPT3L L2	-0.5645
MCQ Accuracy	0.310	0.298	25.94	PWLD NC-L2	-69.4396
				PWLD LIT-L2	50.6198
				ruGPT3S LIT	1.2018
				ruBL L2	-0.8810
FT Time	0.244	0.237	31.36	ruBL L2	0.0461
				nWAS NC-L2	0.2156
				ruGPT3S L2	0.0186
				ruBL LIT	0.9105
MCQ Time	0.182	0.177	32.53	ruBL L2	0.0335
				ruBL LIT	0.0176

Table 3: Multiple regression results

7. Discussion and Conclusion

Addressing our first research question **(RQ1)**, the study reveals the following findings:

1. Non-compositional expression comprehension scores are highest for Belarusian and Ukrainian, languages within the same (East Slavic) group as Russian. The response times for these languages are the lowest.

2. Notably, there is minimal difference in the performance metrics between Belarusian and Ukrainian.

3. Bulgarian, the only representative of the South Slavic group, scored lower than East Slavic languages but higher than West Slavic languages

(Polish and Czech). This could be attributed to the use of Cyrillic script in Bulgarian, which likely facilitated intercomprehension by native Russian speakers.

4. Within the West Slavic group, participants exhibited significantly lower scores in the free translation task for Czech compared to Polish. However, only a slight difference was observed in the multiple-choice question task performance between Czech and Polish.

5. Overall, the observed pattern in scores aligns with the traditional linguistic classification of Slavic languages.

Regarding the second research question **(RQ2)** we demonstrate that:

1. The percentage of correct responses in both experimental tasks exhibits a strong and statistically significant correlation with phonological distance between foreign and Russian non-compositional expressions. For all target languages, this correlation is stronger than the correlation with orthographic distance. Although it may seem surprising, it is in line with previous research (e.g. Jagrova, 2022).

2. Accuracy in MCQ task additionally has a significant positive correlation with the phonological distance between foreign non-compositional and Russian literal expressions, but only for East Slavic languages. The positive correlation suggests that when making a choice between a non-compositional and literal Russian expressions, participants are likely to choose non-compositional expression if the distance between foreign non-compositional and Russian literal expression is large.

3. Response time in both tasks has a stronger relationship with LM surprisal (especially for masked model ruBERTa-large) for all languages except Ukrainian, which supports our initial hypothesis and suggests that advanced language models can reflect the difficulty in congitive processing. We do not observe a strong correlation of response time with any of the linguistic distance variables.

4. Response time for Ukrainian language, in contrast to all other target languages, does not show any significant correlation with LM surprisal. The absence of this correlation suggests a greater difference in the perception of non-compositional expressions between humans and language models in Ukrainian compared to other languages. Additionally, we hypothesize that additional factors, such as cultural influences or variations in participant demographics, may contribute to the observed results for Ukrainian. Further investigation into these potential factors is required to gain a better understanding of this phenomenon.

5. From the multiple regression analysis involving the data for all language sets jointly, we can additionally see the impact of both masked and autoregressive language models on accuracy in both tasks. This fact is significant, considering that the two types account for both the contextual information to the left and the entire sentential context, recognizing their joint importance in predicting the intelligibility scores.

In summary, this research contributes to our understanding of how non-compositional expressions are comprehended across languages, with implications for fields such as linguistics, cognitive science, and natural language processing. Future research could explore the differences of cross-lingual non-compositional comprehension intelligibility in written and spoken modality.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive and insightful feedback on the paper. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102 and by Saarland University (UdS-Internationalisierungsfonds).

Limitations

While this study offers valuable insights into the cross-lingual intelligibility of non-compositional expressions in context, it is essential to acknowledge certain limitations. Firstly, because we rely only on native speakers of Russian as study participants, the findings may not be fully generalizable to other language groups and even to other Slavic languages. Secondly, our analyses were conducted using Language Models specifically tailored for Russian, which means we need to be cautious when applying the results to other languages. Additionally, the predictive factors used in the study, including linguistic distances and surprisal scores, may not fully capture all the complexities of cross-linguistic intelligibility. Factors such as semantic similarity, syntactic structures, and cultural nuances could also play significant roles but were not included in our analysis. Acknowledging and addressing these limitations is crucial for a thorough understanding of the study's findings.

References

Badr M. Abdullah, Marius Mosbach, Iuliia Zaitova, Bernd Möbius, and Dietrich Klakow. 2021. Do Acoustic Word Embeddings Capture Phonological Similarity? An Empirical Study. In *Proceedings of Interspeech 2021*, pages 4194–4198.

Ben Alderson-Day and Charles Fernyhough. 2015. Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, 141(5):931–965.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292. Chapman and Hall/CRC.

Cristina Cacciari and Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of Memory and Language*, 27(6):668–683.

Jelena Golubović Anja Schüppert Femke Swarte Charlotte Gooskens, Vincent J. van Heuven and Stefanie Voigt. 2018. Mutual intelligibility between closely related languages in europe. *International Journal of Multilingualism*, 15(2):169–193.

Kathy Conklin and Norbert Schmitt. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29:82–89.

M. Crocker, V. Demberg, and E. Teich. 2016. Information density and linguistic encoding (ideal). *Künstliche Intelligenz*, 30:77–81.

Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, and Xavier Aumont. 2016. Using phonologically weighted levenshtein distances for the prediction of microscopic intelligibility. In *Annual conference Interspeech (INTERSPEECH 2016)*, page 650.

Charlotte Gooskens. 2013. Experimental methods for measuring intelligibility of closely related language varieties. In *The Oxford Handbook of Sociolinguistics*.

Charlotte Gooskens and Femke Swarte. 2017. Linguistic and extra-linguistic predictors of mutual intelligibility between germanic languages. *Nordic Journal of Linguistics*, 40:123–147.

Charlotte Gooskens and Vincent Van Heuven. 2022. *Mutual intelligibility*, pages 51–95. Cambridge University Press.

Leonid Iomdin. 2015. Microsyntactic constructions formed by the Russian word raz. *SLAVIA cˇasopis pro slovanskou filologii*, 84(3).

Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press UK.

Klara Jagrova. 2022. *Reading Polish with Czech Eyes: Distance and Surprisal in Quantitative, Qualitative, and Error Analyses of Intelligibility*. universaar.

Klara Jagrova, Tania Avgustinova, Irina Stenger, and Andrea Fischer. 2018. Language models, surprisal and fantasy in Slavic intercomprehension. *Computer Speech Language*, 53.

Klára Jágrová, Michael Hedderich, Marius Mosbach, Tania Avgustinova, and Dietrich Klakow. 2021. On the correlation of context-aware language models with the intelligibility of polish target words to czech readers. *Frontiers in Psychology*, 12.

Jacek Kudera, Irina Stenger, Philip Georgis, Bernd Möbius, Tania Avgustinova, and Dietrich Klakow. 2023. Cross-linguistic intelligibility of idiomatic phrases in polish-russian translation tasks. In Jean-Pierre Colson, editor, *Phraseology, Constructions and Translation: Corpus-based, Computational and Cultural Aspects*, pages 237–249. Presses Universitaires de Louvain.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Tomáš Machálek. 2020. Kontext: Advanced and flexible corpus query interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Marius Mosbach, Irina Stenger, Tania Avgustinova, and Dietrich Klakow. 2019. incom.py - a toolbox for calculating linguistic distances and asymmetries between related languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 810–818, Varna, Bulgaria. INCOMA Ltd.

Robert Möller and Ludger Zeevaert. 2015. Investigating word recognition in intercomprehension: Methods and findings. *Linguistics*, 53.

S. B. Needleman and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453.

B.H. Partee. 2008. *Compositionality in Formal Semantics: Selected Papers*. Explorations in Semantics. Wiley.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report.

Giulia Rambelli, Emmanuele Chersoni, Marco S. G. Senaldi, Philippe Blache, and Alessandro Lenci. 2023. Are frequent phrases directly retrieved like idioms? an investigation with self-paced reading and language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 87–98, Dubrovnik, Croatia. Association for Computational Linguistics.

SberDevices. 2023. ruroberta-large. Hugging Face Model Hub.

Anja Schüppert, Johannes C. Ziegler, Holger Juul, and Charlotte Gooskens. 2022. On-line activation of l1 danish orthography enhances spoken word recognition of swedish. *Nordic Journal of Linguistics*, 45:80–98.

Anna Siyanova-Chanturia, Kathy Conklin, and Norbert Schmitt. 2011. Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, 27(2):251–272.

I. Stenger and T. Avgustinova. 2021. On slavic cognate recognition in context. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference 'Dialogue'*, volume 20, pages 660–668, Moscow, Russia.

Irina Stenger. 2019. *Doctoral Dissertation: Zur Rolle der Orthographie in der slavischen Interkomprehension mit besonderem Fokus auf die kyrillische Schrift*. Ph.D. thesis, Saarbrücken: universaar.

Irina Stenger, Philip Georgis, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. 2022. Modeling the impact of syntactic distance and surprisal on cross-Slavic text comprehension. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7368–7376, Marseille, France. European Language Resources Association.

Irina Stenger, Klara Jagrova, and Tania Avgustinova. 2020. The INCOMSLAV platform: Experimental website with integrated methods for measuring linguistic distances and asymmetries in receptive multilingualism. In *Proceedings of the LREC 2020 Workshop on "Citizen Linguistics in Language Resource Development"*, pages 40–48, Marseille, France. European Language Resources Association.

Irina Stenger, Klára Jágrová, Andrea Fischer, Tania Avgustinova, Dietrich Klakow, and Roland Marti. 2017. Modeling the impact of orthographic coding on czech–polish and bulgarian–russian reading intercomprehension. *Nordic Journal of Linguistics*, 40(2):175–199.

Roland Sussex and Paul Cubberley. 2006. *The Slavic Languages*. Cambridge University Press, Cambridge.

Debra Titone, Kyle Lovseth, Kristina Kasparian, and Mehrgol Tiv. 2019. Are figurative interpretations of idioms directly retrieved, compositionally built, or both? evidence from eye movement measures of reading. *PsyArXiv*.

Jan Vanhove and Raphael Berthele. 2015. Item-related determinants of cognate guessing in multilinguals. *Crosslinguistic Influence and Crosslinguistic Interaction in Multilingual Language Learning*, 95:118.

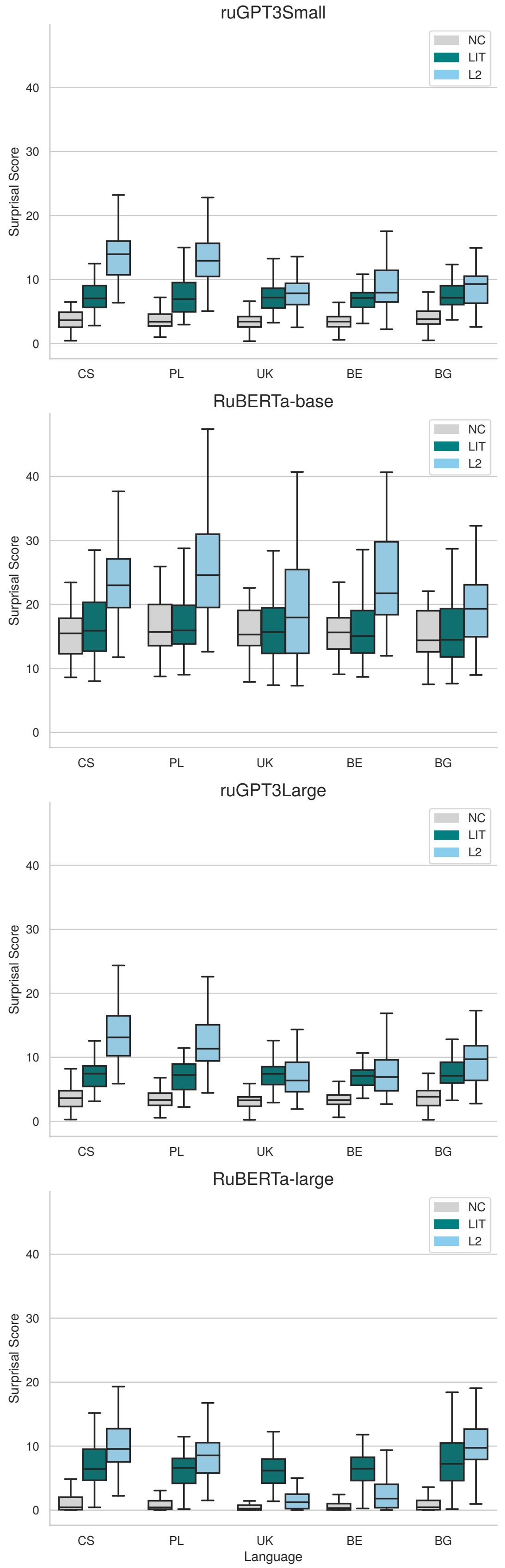
Francesco Vespignani, Paolo Canal, Nicola Molinaro, Sergio Fonda, and Cristina Cacciari. 2009. Predictive mechanisms in idiom comprehension. *Journal of cognitive neuroscience*, 22:1682–700.

Iuliia Zaitova, Badr Abdullah, and Dietrich Klakow. 2022. Mapping phonology to semantics: A computational model of cross-lingual spoken-word recognition. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Iuliia Zaitova, Irina Stenger, and Tania Avgustinova. 2023. Microsyntactic unit detection using word embedding models: Experiments on slavic languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1265–1273. INCOMA Ltd.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion.

A. Language Model Surprisal



Surprisal across models and languages for LIT, NC, and L2 phrases

B. Correlation Tables

Correlation of distances and surprisal with FT task accuracy

Metrics	BE	BG	CS	PL	UK
nWAS					
NC-L2	-0.206 (NS)	-0.287*	-0.025 (NS)	0.067 (NS)	-0.216 (NS)
PWLD					
NC-L2	**-0.405****	**-0.471*****	**-0.361****	**-0.428*****	**-0.606*****
ruGPT3S					
NC	-0.297*	0.035 (NS)	-0.082 (NS)	0.011 (NS)	-0.180 (NS)
L2	-0.296*	-0.232 (NS)	-0.236 (NS)	-0.162 (NS)	-0.131 (NS)
ruGPT3L					
NC	-0.303*	0.053 (NS)	-0.046 (NS)	0.050 (NS)	-0.151 (NS)
L2	-0.299*	-0.247 (NS)	-0.257*	-0.220 (NS)	-0.106 (NS)
ruBL					
NC	0.093 (NS)	0.046 (NS)	-0.097 (NS)	0.061 (NS)	-0.286*
L2	-0.062 (NS)	-0.173 (NS)	-0.233 (NS)	-0.209 (NS)	-0.255*

Correlation of distances and surprisal with MCQ task accuracy

Metrics	BE	BG	CS	PL	UK
nWAS					
NC-L2	-0.102 (NS)	-0.266*	-0.171 (NS)	-0.001 (NS)	-0.279*
LIT-L2	-0.044 (NS)	-0.018 (NS)	0.064 (NS)	-0.097 (NS)	0.064 (NS)
PWLD					
NC-L2	-0.229 (NS)	**-0.417****	**-0.283***	**-0.385****	**-0.502*****
LIT-L2	**0.429*****	-0.035 (NS)	0.063 (NS)	0.056 (NS)	0.307*
ruGPT3S					
NC	-0.216 (NS)	-0.147 (NS)	0.098 (NS)	0.084 (NS)	-0.030 (NS)
LIT	-0.015 (NS)	0.055 (NS)	0.143 (NS)	0.301*	0.254 (NS)
L2	-0.021 (NS)	-0.064 (NS)	0.210 (NS)	-0.019 (NS)	0.001 (NS)
ruGPT3L					
NC	-0.208 (NS)	-0.124 (NS)	0.115 (NS)	0.007 (NS)	0.047 (NS)
LIT	0.018 (NS)	0.039 (NS)	0.117 (NS)	0.260*	0.253 (NS)
L2	-0.030 (NS)	-0.025 (NS)	0.177 (NS)	-0.045 (NS)	0.076 (NS)
ruBL					
NC	0.041 (NS)	-0.025 (NS)	-0.013 (NS)	-0.063 (NS)	-0.169 (NS)
LIT	0.195 (NS)	0.188 (NS)	0.122 (NS)	0.338**	0.153 (NS)
L2	0.040 (NS)	-0.042 (NS)	0.189 (NS)	-0.010 (NS)	-0.129 (NS)

*=p$< .05$, **=p$< .01$, and ***=p$< .001$. Pearson correlation of intelligibility metrics

Correlation of distances and surprisal with for FT task time

Metrics	BE	BG	CS	PL	UK
nWAS					
NC-L2	0.145 (NS)	0.078 (NS)	0.312*	0.133 (NS)	-0.039 (NS)
PWLD					
NC-L2	0.060 (NS)	-0.047 (NS)	0.026 (NS)	0.221 (NS)	0.068 (NS)
ruGPT3S					
NC	0.225 (NS)	0.003 (NS)	0.278*	-0.028 (NS)	0.143 (NS)
L2	0.363**	**0.318***	0.501***	0.201 (NS)	0.177 (NS)
ruGPT3L					
NC	0.265*	0.009 (NS)	0.280*	0.020 (NS)	0.080 (NS)
L2	0.410**	0.277*	0.441***	0.182 (NS)	0.209 (NS)
ruBL					
NC	0.311*	-0.120 (NS)	0.223 (NS)	0.149 (NS)	**0.222** (NS)
L2	**0.443*****	0.135 (NS)	**0.547*****	**0.304***	0.217 (NS)

Correlation of distances and surprisal with MCQ task time

Metrics	BE	BG	CS	PL	UK
nWAS					
NC-L2	0.118 (NS)	0.078 (NS)	0.285*	0.023 (NS)	-0.101 (NS)
LIT-L2	-0.144 (NS)	**-0.322***	0.125 (NS)	0.007 (NS)	-0.180 (NS)
PWLD					
NC-L2	0.000 (NS)	-0.019 (NS)	0.040 (NS)	0.135 (NS)	0.062 (NS)
LIT-L2	0.007 (NS)	-0.220 (NS)	0.151 (NS)	0.057 (NS)	-0.013 (NS)
ruGPT3S					
NC	0.229 (NS)	0.015 (NS)	0.213 (NS)	-0.093 (NS)	0.143 (NS)
LIT	0.155 (NS)	-0.008 (NS)	0.167 (NS)	0.235 (NS)	0.013 (NS)
L2	0.311*	**0.322***	0.374**	0.089 (NS)	0.131 (NS)
ruGPT3L					
NC	0.278*	0.029 (NS)	0.201 (NS)	-0.047 (NS)	0.113 (NS)
LIT	0.183 (NS)	-0.032 (NS)	0.199 (NS)	0.254 (NS)	0.019 (NS)
L2	0.368**	0.289*	0.309*	0.079 (NS)	0.174 (NS)
ruBL					
NC	0.330*	-0.102 (NS)	0.191 (NS)	0.154 (NS)	0.083 (NS)
LIT	0.142 (NS)	0.200 (NS)	0.323 (NS)	0.288*	-0.080 (NS)
L2	**0.457*****	0.102 (NS)	**0.452*****	**0.308***	**0.215** (NS)

*=p$< .05$, **=p$< .01$, and ***=p$< .001$. Pearson correlation of time metrics

Using a Language Model to Unravel Semantic Development in Children's Use of a Dutch Perception Verb

Bram van Dijk[1], Max van Duijn[1], Li Kloostra[2], Marco Spruit[1], and Barend Beekhuizen[3]
[1]Leiden Institute of Advanced Computer Science, [2]Utrecht University, [3]University of Toronto
Corresponding author: b.m.a.van.dijk@liacs.leidenuniv.nl

Abstract

We employ a Language Model (LM) to gain insight into how complex semantics of Dutch Perception Verb (PV) *zien* ('to see') emerge in children. Using a Dutch LM as representation of mature language use, we find that for ages 4-12y 1) the LM accurately predicts PV use in children's freely-told narratives; 2) children's PV use is close to mature use; and 3) complex PV meanings with attentional and cognitive aspects can be found. Our approach illustrates how LMs can be meaningfully employed in studying language development, hence takes a constructive position in the debate on the relevance of LMs in this context.

Keywords: language development, language models, computational modelling, semantics, pragmatics

1. Introduction

Recent Language Models (LMs) based on Transformer architectures (Vaswani et al., 2017) reflect semantic knowledge present in a language community. BERT vectors (Devlin et al., 2019), for example, are able to distinguish different senses of the same word (Rogers et al., 2020; Vulić et al., 2020; Wiedemann et al., 2019). These LMs implement the distributional hypothesis that words with similar meanings tend to occur in similar contexts, and they represent both word type and word token meanings with real-valued vectors (Lenci and Sahlgren, 2023). The latter allows LMs to encode polysemy and different usages of words.

Despite this, LMs' relevance in the context of language development is disputed: their architecture and volume of training input have been argued to make them incomparable to children (e.g. Bunzeck and Zarrieß, 2023; Prystawski et al., 2022; Warstadt and Bowman, 2022). Yet, others argue that LMs can show which linguistic phenomena are *in principle* learnable from distributional information, bearing on learnability debates (Contreras Kallens et al., 2023; Piantadosi, 2023; Wilcox et al., 2023).

Here we leverage LMs' rich semantic information to gain insight in children's semantic and pragmatic development. Addressing the question whether children's pragmatic use of lexical items develops over time or, conversely, is adult-like from the start, we use a Dutch LM as representation of mature language use and study the Dutch Perception Verb (PV) *zien* ('to see'). We find that children's use of *see* is close to mature use across the 4-12y age range, and that for all ages the familiar mature usage patterns of the verb can be identified. As such, the paper further illustrates the relevance of LMs in studying language development, by reflecting on LMs as representations of mature language use and setting up appropriate tasks and metrics.

2. Background

Little empirical work employs modern LMs in language development, the exception being work comparing word acquisition in children and LMs (Chang and Bergen, 2022; Laverghetta and Licato, 2021). This is understandable given the debate on the validity of LMs in the child context: LMs and children differ in key respects including word exposure (Warstadt and Bowman, 2022) and learning mechanisms (Bunzeck and Zarrieß, 2023).

Still, LMs are arguably useful representations of *mature language use* by being trained on corpora of adult language, and are therefore of value in modelling language understanding. LMs can be viewed as an incremental methodological step compared to earlier corpus studies comparing children's verb use to mature use, that relied on manual annotation or feature engineering to identify different senses of mature verb use (e.g. Adricula and Narasimhan, 2009; Parisien and Stevenson, 2009), but different senses, as we will show, can also be conveniently retrieved from LLMs. These and other considerations have led to increasing acknowledgement of LMs' relevance for analysing language development (Contreras Kallens et al., 2023; Lappin, 2023), and efforts to make LMs more comparable to the child context (Warstadt et al., 2023).

Here we address the relevance of LMs in the developmental context by analysing children's lexical semantic development with LMs. We target children's use of Dutch PV *zien* ('to see') as a case study, which has been frequently analysed in language development (e.g. Davis, 2020; Davis and Landau, 2021). Studies of perception verbs across languages have shown that visual perception verbs have extended meanings beyond their denotational meaning 'entity X visually perceives object or event Y', that involve additional aspects of e.g. *attention* ('Let's see if I can find the keys') and *cognition* ('I

The Workshop on Cognitive Aspects of the Lexicon, pages 98–106
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

see what you mean') (San Roque et al., 2018; San Roque and Schieffelin, 2019). Such meaning extensions are salient for children with a limited lexicon, where meaning extension of known words allows children to express new meanings efficiently (Nerlich and Clarke, 1999). In addition, since visual perception is argued to have strong metaphorical mappings to knowledge and understanding (e.g. Johnson, 1999), *see* can be a window onto how children learn to represent (socio-)cognitive content with language (Sweetser, 1990).

This work addresses the question when meaning extension occurs. Some argue that literal understandings of PVs emerge first in young children (e.g. Davis, 2020; Davis and Landau, 2021; Elli et al., 2021; Landau and Gleitman, 2009), while others argue pragmatic meanings are likely present early due to the social situatedness of language learning (e.g. Enfield, 2023; San Roque and Schieffelin, 2019). In the latter case, the discursive relation between the visual perception event and the events surrounding it may be more salient for a language learner than the encoding of visual perception per se. For example, a young child's utterance *see ball* may be followed by the caregiver showing the ball, or focusing its attention on the ball — further attentional aspects that are likely relevant components of the message for the child beyond the denotational content of visual perception having taken place. While focusing on a single verb may seem limited, we believe as a case study, visual perception verbs are well-chosen as a starting point for generalising the proposed approach, since their acquisitional pathway and pragmatic usages (as described above) are well understood.

We focus on children's use of *see* in ChiSCor, a corpus of freely-told stories by Dutch children (4-12y) in classroom settings (van Dijk et al., 2023b), since complex PV meanings can be especially relevant in the narrative domain. For example, that character X *sees* entity Y may not only imply that X literally perceives Y, but also that X *evaluates* Y or *discovers* Y. Such information, which may be crucial for the 'tellability' of the story (Labov and Waletzky, 1967), can be efficiently transmitted through PVs. Narratives are 'natural' sandboxes for children to challenge their language competence in various ways (Frizelle et al., 2018), including the development of lexical pragmatics.

3. Methods

Language data – We extracted all 308 occurrences of *see* from 619 stories of 442 children (4-12y) in ChiSCor. We manually inspected these occurrences and removed unintelligible usages (mainly transcription errors) as well as stories exceeding a context window larger than 512 tokens, resulting in 210 occurrences. We assigned occurrences to a Young (4-6), Middle (6-9) or Old (9-12) age group, following the age binning in Dutch primary education, and included only PV occurrences from one story per child, resulting in 30 Young, 82 Middle and 42 Old PV occurrences. To balance the sample across age groups, we randomly sampled 30 occurrences from the Middle and Old age group.

A known problem with LMs is that data contamination can lead them to solve tasks by memorisation (Deng et al., 2023). ChiSCor is likely not in the train data of recent LMs, as the corpus is recent and 'hidden' behind view-only links in research papers. Further, ChiSCor's free storytelling is unlike other available Dutch corpora that involve language elicitation and as such constitutes language that tests LMs' generalisation capabilities.

LMs as benchmark models – Using LMs as representation of mature language use requires evidence that the LM models the linguistic phenomenon and domain at issue reliably. We draw on findings that word representations in BERT encode rich semantic information about word polysemy (Garí Soler and Apidianaki, 2021; Wiedemann et al., 2019), although not perfectly. Also, Dutch LMs are for a large part trained on narrative texts (e.g. De Vries et al., 2019; Delobelle et al., 2020), and LMs in general have been shown to model coherence in written narratives well (Laban et al., 2021). In sum, earlier work supports the idea that LMs encode mature PV use in narratives.

Choice of LMs – For reasons of computational efficiency, validity with respect to the child context, and reproducibility, we chose `RobBERT-2023-dutch-large`, a Dutch BERT-like LM (Delobelle et al., 2020). RobBERT has 455M parameters trained on 19.5B tokens and is more in line with the 100M training input a 10-year-old has seen (Warstadt and Bowman, 2022), compared to often employed large LMs like GPT-3 (175B parameters, 500B tokens (Brown et al., 2020)). RobBERT is accessible through the `HuggingFace Transformers` ecosystem (Wolf et al., 2019).

Recent work on making LMs relevant to human language acquisition in the BabyLM challenge (Warstadt et al., 2023), highlighted smaller LMs with optimised architectures and train objectives, and curated train data for training developmentally plausible models (Samuel et al., 2023). However, such Dutch LMs are not yet available and training models from scratch is generally not feasible for researchers studying language acquisition. RobBERT was a fitting resource as it is optimised compared to BERT and has a simpler training objective (masked language modelling only) (Liu et al., 2019). These aspects go some way towards the findings of the BabyLM challenge (Samuel et al., 2023; Warstadt et al., 2023).

Task design and metrics – To use LMs as representations of mature language use, zero-shot evaluation settings as described by Laban et al. (2021) are preferred. This means using LMs off-the-shelf without further pre-training on the target domain or fine-tuning to stay close to the mature language use encoded in the LM, similar to how factual knowledge can be retrieved from LMs without fine-tuning (Petroni et al., 2019).

We use various possibilities available through LMs to assess whether and how children's use of *see* differs from mature use.

Our first task consists of predicting *see* in children's narratives. We present RobBERT with stories containing a masked instance of *see*, as in the (translated) excerpt in (1):

(1) [...] one time robot was travelling. and all of a sudden he **<mask>** a wolf. and he ran away quickly. [...] (Story ID 052301)

In our experiment we provided full stories as context to RobBERT, which varied in number of words ($\bar{x} = 187, \sigma = 108$). If children's usage differs from adults, the LM might have difficulty predicting the PV correctly.

As a second measure, we compute the negative log-likelihood NLL or surprisal for a prediction for a masked token w_m with $NLL(w_m) = -\log p(w_m | w_{1...m-1}, w_{m+1...n})$ with the `fill-mask` pipeline from `HuggingFace Transformers`. This measure provides further context to the predictive accuracy measure presented above: lower NLL implies that the predicted token is less surprising i.e. closer to mature use as encoded in the LM, and more generally indicates how well a given context supports a specific token on the masked position (PV or other).

Lastly, we use the tokens in RobBERT's top-5 predictions for masked instances of *see* as 'near neighbours' that can reveal the additional discursive meanings that the usage of PVs support. Our data and notebooks are available at `https://shorturl.at/jquVX`.

4. Results

Predictive accuracy – First, we assessed RobBERT's overall performance in predicting *see* at masked positions in all 90 PV occurrences. Accuracy is overall high (.83, Table 1), and although lower for Young (.70) we found no significant difference in accuracy between ages with an ANOVA ($F_{2,87} = 2.974, p = .056$). This shows that RobBERT models children's PV use in the narrative domain well. The 15 errors were mainly in Young and showed confusion of *seeing* with 'finding', 'having', 'looking' and 'getting', meaning that contexts under-constrained the use of *see*. Although these other verbs can be valid tokens on masked positions (e.g. 'found' in (1)), here our aim was to see if RobBERT adequately models that *see* can subsume such other possible meanings in narratives.

Metric	Young	Middle	Old	Overall
Accuracy	.70 (30)	.90 (30)	.90 (30)	.83 (90)
Surprisal	.40 (21)	.23 (27)	.32 (27)	.31 (75)
Top-5	1.00 (30)	1.00 (30)	.97 (30)	.99 (90)

Table 1: Metrics for RobBERT. Accuracy: percentage that *see* was predicted. Surprisal: NLL computed for predictions of *see*. Top-5: proportion that *see* was in top-5 predictions. Number of PV occurrences (i.e. observations) in parentheses.

Surprisal – Second, we analysed potential age effects in mean surprisal for 75 correct predictions of *see*. For example, RobBERT may be less surprised by PV use for Old compared to Young or Middle, indicating PV use of Old children is closer to mature use than Young. Interestingly, surprisal distributions are close to 0 for all ages (Figure 1), and although mean surprisal between Young, Middle, and Old differs (Table 1), pairwise comparisons with Tukey's HSD (Tukey, 1949) revealed no significant age effects. This suggests that PV use by children of all ages is about equally close to mature use.

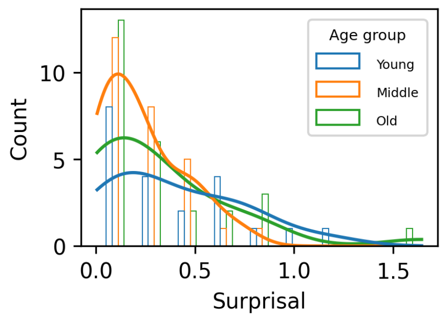


Figure 1: Surprisal distributions.

Top-5 alternative predictions – For virtually all age groups, *see* is in the top-5 predictions (Table 1), which supports the idea that by examining top-5s we get insight in extended meanings of *see*. For 90 PV occurrences and their top-5s (450 tokens) we lemmatised tokens and removed *see* and lemmas that were not verbs (e.g. 'many', 'and', 'at'), resulting in 304 lemmas. We then took the set and classified 65 lemmas as having roughly 'external', 'internal', or 'other' meaning. External implies a meaning pertaining to plain action (e.g. 'to go', 'to come', 'to carry', 'to throw'); internal a meaning pertaining to an attentional (e.g. 'to notice', 'to meet') or cognitive state (e.g. 'to think', 'to know'). Other pertains to auxiliary verbs and PVs not the focus of the current study (e.g. 'to have', 'to hear'). The

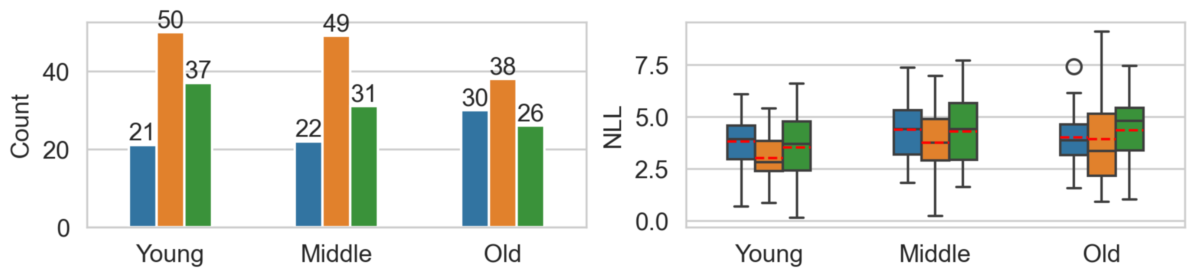


Figure 2: Frequencies (left) and surprisal dist. (right) of internal (blue), external (orange) and other (green) meanings of 304 top-5 lemmas. Bars (left) stack to 100%; dashed red lines (right) indicate means.

Age	Ex.	PV context
Young	(2)	.. and when he returned. then he saw/knew that the princess was gone. and they lived happily ever after. (102901)
	(3)	.. and then they were lost again. and then they saw/searched the castle. and then they went in the castle. (122901)
	(4)	.. but then the teacher came and then she was already too late. the teacher had seen/caught them. and then you get a punishment from the teacher. (033401)
Middle	(5)	.. but then they lost each other all of a sudden. and then Wergje saw/met another rabbit. and it asked how are you called. (072301)
	(6&7)	.. because when he was home. then he saw/noticed/discovered that he had the other scales. but then he went to fly on it and he wanted to find his own dragon again. (022301)
Old	(8)	.. once arrived at the cave Puta completely forgot that you were not allowed to touch the big diamond. Puta saw/checked out the diamond and found it so beautiful. and he touched it accidentally. (034801)
	(9)	.. so then the fat little king went on his fat broom to the cry for help. and what did he see/think. the cry came from a little fat guinea pig that looked very much like the king. (023801)
	(10)	.. and he ever wanted one time to try it with his eyes closed. to see/test can I grab that donut well with my eyes closed. (034501)

Table 2: Translated PV contexts with top-5 internal lemmas (underlined) with lowest surprisal. Story IDs given in parentheses. All excerpts were translated by the first author.

idea is that top-5 lemmas indicate what possible meanings PV contexts support, even if these lemmas are not necessarily intuitive substitutions. For example, substituting ‘threw’ for **<mask>** in (1) renders the excerpt less intuitive. Yet, this immediate context as a sequence of *external* actions better supports understanding *seeing* also as a causal part of a sequence of external actions, than as *seeing* as part of narrative components reflecting a character’s attentional or cognitive *internal* states (cf. examples in Table 2).

We assessed frequencies of external, internal and other meanings, and their mean surprisal over age groups to identify potential age differences in occurrence and closeness to mature use. Regarding frequency, although external and other meanings decrease over age while internal meanings increase over age (Figure 2, left), we found no significant age effects with a χ^2 test $\chi^2(4, N = 304) = 5.044, p = .283$, suggesting that all the different meanings are about equally frequent in Young, Middle and Old groups. Regarding surprisal (Figure 2, right), distributions for external, internal and other meanings are relatively similar both within and between age groups. Pairwise comparisons with Tukey’s HSD found only a significant difference at the $p < .05$ level between mean surprisal for external meanings for Young and Old.

We illustrate complex meanings of *see* present in all age groups, by providing the three internal meanings that were closest to mature use (i.e. with lowest surprisal) and their PV contexts in Table 2. We make three observations. First, internal meanings with attentional and cognitive aspects can be but are not exclusively cued by surface linguistic frames such as complementation that RobBERT simply picks up, as example (4) and (9) show. In (4) ‘caught’ implies that the teacher knows what the ‘she’ character is up to; in (9) ‘think’ renders the realisation where the cry of help is coming from a representation in the mind of the king. Second, internal meanings are varied: from more purely attentional where characters simply become aware of something or find something out as in (6&7), to more social (5), and evaluative attentional aspects (8). Third, although internal meanings with cognitive aspects have the most abstract lemmas (‘think’, ‘know’) that are argued to be harder to master (Barak et al., 2012), cognitive meanings were found in both Young (2), (4) and Old (9) children.

5. Discussion

Our results show that complex meanings of the Dutch perception verb *zien* ('to see') are about equally frequent in all age groups and that children's use of the PV is overall not significantly different from mature use. This contrasts with earlier work that has argued that children initially acquire more literal meanings of PVs (Adricula and Narasimhan, 2009; Davis, 2020; Davis and Landau, 2021; Elli et al., 2021; Landau and Gleitman, 2009) (Section 2), although we note that children in our sample are older (four years and older) than children in earlier studies (typically between two and four years).

Our result aligns with the idea that it is the social context that cues various complex senses of *see* in children (e.g. Enfield, 2023; San Roque and Schieffelin, 2019), and with the idea that (young) children may employ PVs like *see* as linguistic devices for learning to represent cognitive and attentional states (Johnson, 1999; Sweetser, 1990). We argue that our finding can be explained by the social context provided by live storytelling. PVs like *see* are linguistic devices for efficiently communicating about characters' attentional and cognitive states that are key to understanding the story, as PVs can compress redundant information that would make the story tedious. Earlier work has shown that, in children's live storytelling, contexts of PV like *hear* and *see* are coherent and clear, as evidenced by the rich PV vectors that can be trained from limited amounts of narrative data (van Dijk et al., 2023b).

Narrative language data may explain the contrast between our and earlier findings as storytelling has been argued to solicit 'maximal behaviour' in that it challenges children's linguistic competence (Frizelle et al., 2018; Southwood and Russell, 2004), more than the speech produced by children in child-caregiver interactions would do, which typically take place in mundane contexts. Some earlier work contrasting with our results relied on language data from such child-caregiver interactions (e.g. Davis and Landau, 2021; Adricula and Narasimhan, 2009). The latter work also employed smaller sample sizes with less unique children and more PV use per child compared to the current study, which compresses the variation in complex semantics we find in our analysis.

Interestingly, RobBERT accurately predicted *see* in narratives of children of all ages; we argue that this is not a mere frequency effect (i.e. *see* being more frequent in train data than alternatives), given that top-5 predictions often reveal RobBERT's correct mapping of the nuanced senses of PVs. Also, RobBERT's aptitude in handling PV use in narratives is interesting insofar children's stories are not obvious regarding wording, characters and themes. One issue pointed out by a reviewer is whether LMs with Transformer architectures are the best fit for representing linguistic knowledge of a mature Dutch language user, or whether other models should be used, e.g. from the BabyLM challenge (Warstadt et al., 2023). The best-performing LMs in this challenge employed Transformer architectures that are essentially optimised versions of vanilla BERT models regarding training objective, architecture and dataset (Samuel et al., 2023). With our choice for RobBERT we aimed to make the comparison to the human case as valid as possible with an existing resource (see Section 3).

In any case, from the BabyLM challenge we learn that the Transformer architecture is also in more modest training setups a powerful encoder of linguistic information. Our claim is not that Transformers are therefore good (cognitive) models of human language users, which is still debated (e.g. Paape, 2023; van Dijk et al., 2023a). Rather, when it comes to specific linguistic aspects such as mature semantic and pragmatic knowledge, LMs as sophisticated distributional learners represent this information in a convenient fashion. For using such computational models as representations of mature language use, the primary question is if their *behaviour* for a specific linguistic phenomenon is sufficiently complex, which for many modern BERT-like models seems the case. But representations of mature use could also be created in other ways, e.g. by clustering different verb senses with features based on verb argument structure in a large corpus of mature language use. Thus, LMs are more of an analytical tool here than direct models of humans. That said, it is still worthwhile and necessary to make LMs more similar to the human context.

6. Conclusion

This paper provided a case study on Dutch children's (4-12y) use of *zien* ('to see') and the emergence of complex semantics in this perception verb. We showed that 1) a recent Dutch LM can predict use of *see* in narratives for different ages reliably; 2) children's use of *see* is close to mature use for all ages; and 3) complex meanings of *see* with attentional and cognitive aspects can be found across all ages. Our results align with work that argues that meaning extension occurs early in children and with the idea that via perception verbs, children may learn to represent socio-cognitive content.

We also showed how LMs can be meaningfully leveraged in developmental contexts. We hope to provide future researchers with useful reflection on how to proceed when using LMs as representations of mature language use, choosing models, and setting up tasks and metrics.

7. Limitations and Ethical Considerations

A limitation of this study is that we provided the whole story as context for predicting a masked occurrence of *zien* ('to see'), but for space limitations only could discuss complex meanings with smaller story excerpts as in Table 2. This may suggest that complex PV meanings can be determined from small pieces of narrative after all. Yet, when doing the same task with smaller PV contexts as in Table 2, i.e. a sentence before and after the sentence featuring an occurrence of *see*, RobBERT's overall accuracy drops from .83 to .57 and overall surprisal increases from .31 to .59, (see Table 1) which suggests that RobBERT needs to take the whole story into account to model PV use adequately. This means that there is more relevant information in the context beyond what we show in the immediate PV context that render RobBERT's predictions of masked tokens accurate and support additional meanings of *see*.

Another limitation is that we had to translate story excerpts to English, as also providing Dutch excerpts required too much space. Some awkwardness in translations could not be avoided. For example, Dutch has a verb 'betrappen' that always has a cognitive meaning similar to 'catching somebody red-handed', whereas 'catching' in English can also have a more obvious action-related meaning. 'Betrappen' was a token prediction in RobBERT's top-5 with low surprisal that we had to translate as 'caught' in example (4) in Table 2.

In this study we used the ChiSCor story corpus and we refer to van Dijk et al. (2023b) for further details regarding ethical considerations and approval that was obtained for collecting language data from children. Regarding computational efficiency, we chose a relatively small, open and free to use language model that can also be employed with limited computational resources.

8. Acknowledgements

This research was done in the context of Max van Duijn's research project 'A Telling Story' (with project number VI.Veni.191C.051), which is financed by the Dutch Research Council (NWO). The collaboration with Barend Beekhuizen from the University of Toronto was made possible by funding from The Leiden University/Swaantje Mondt Fund (reference number W232234-1-097).

9. Bibliographical References

Norielle Adricula and Bhuvana Narasimhan. 2009. 'understanding is understanding by seeing': Visual perception verbs in child language. In *Proceedings of the 44th Boston University Conference on Language Development*, pages 18–27, Boston. Somerville, MA: Cascadilla Press.

Libby Barak, Afsaneh Fazly, and Suzanne Stevenson. 2012. Modeling the Acquisition of Mental State Verbs. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 1–10, Montréal, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bastian Bunzeck and Sina Zarrieß. 2023. GPT-wee: How Small Can a Small Language Model Really Get? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 35–46, Singapore. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word Acquisition in Neural Language Models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Pablo Contreras Kallens, Ross Deans Kristensen-McLachlan, and Morten H. Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256.

E. Emory Davis. 2020. *Does seeing mean believing? The development of children's semantic representations for perception verbs*. Ph.D. thesis, The Johns Hopkins University.

E. Emory Davis and Barbara Landau. 2021. Seeing and believing: the relationship between perception and mental verbs in acquisition. *Language Learning and Development*, 17(1):26–47.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle and François Remy. 2023. RobBERT-2023: Keeping Dutch Language Models Up-To-Date at a Lower Cost Thanks to Model Conversion. *The 33rd Meeting of Computational Linguistics in The Netherlands (CLIN 33)*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*.

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.

Giulia V. Elli, Marina Bedny, and Barbara Landau. 2021. How does a blind person see? developmental change in applying visual verbs to agents with disabilities. *Cognition*, 212:104683.

Nick J. Enfield. 2023. Linguistic concepts are self-generating choice architectures. *Philosophical Transactions of the Royal Society B*, 378(1870):20210352.

Rita E. Frank and William S. Hall. 1991. Polysemy and the acquisition of the cognitive internal state lexicon. *Journal of Psycholinguistic Research*, 20(4):283–304.

Pauline Frizelle, Paul A. Thompson, David McDonald, and Dorothy V.M. Bishop. 2018. Growth in syntactic complexity between four years and adulthood: Evidence from a narrative task. *Journal of Child Language*, 45(5):1174–1197.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Christopher Johnson. 1999. Metaphor vs. Conflation in the Acquisition of Polysemy: The Case of SEE. In Masako K. Hiraga, Sherman Wilcox, and Chris Sinha, editors, *Cultural, Psychological and Typological Issues in Cognitive Linguistics: Selected papers of the bi-annual ICLA meeting in Albuquerque*, pages 155–169.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. Can Transformer Models Measure Coherence In Text: Re-Thinking the Shuffle Test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics.

William Labov and Joshua Waletzky. 1967. Narrative analysis: Oral versions of personal experience. In J Holm, editor, *Essays on the Verbal and Visual Arts*, pages 12–44. University of Washington Press.

Barbara Landau and Lila R. Gleitman. 2009. *Language and experience: Evidence from the blind child*. Harvard University Press.

Shalom Lappin. 2023. Assessing the Strengths and Weaknesses of Large Language Models. *Journal of Logic, Language and Information*, pages 1–12.

Antonio Laverghetta and John Licato. 2021. Modeling age of acquisition norms using transformer networks. In *The International FLAIRS Conference Proceedings*, volume 34.

Alessandro Lenci and Magnus Sahlgren. 2023. *Distributional semantics*. Cambridge University Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A roustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Brigitte Nerlich and David D. Clarke. 1999. Elements for an integral theory of semantic change and semantic development. In *Meaning Change–Meaning Variation. Workshop held at Konstanz*, volume 1, pages 123–134.

Dario Paape. 2023. When Transformer models are more compositional than humans: The case of the depth charge illusion. *Experiments in Linguistic Meaning*, 2:202–218.

Christopher Parisien and Suzanne Stevenson. 2009. Modelling the acquisition of verb polysemy in children. In *Proceedings of the CogSci2009 Workshop on Distributional Semantics beyond Concrete Concepts*, pages 17–22, Austin, Texas. Cognitive Science Society.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Steven Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Steven T. Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Ben Prystawski, Erin Grant, Aida Nematzadeh, Spike W.S. Lee, Suzanne Stevenson, and Yang Xu. 2022. The emergence of gender associations in child language development. *Cognitive Science*, 46(6):e13146.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Magnus Sahlgren and Fredrik Carlsson. 2021. The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point. *Frontiers in Artificial Intelligence*, 4:682578.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Lila San Roque, Kobin H. Kendrick, Elisabeth Norcliffe, and Asifa Majid. 2018. Universal meaning extensions of perception verbs are grounded in interaction. *Cognitive Linguistics*, 29(3):371–406.

Lila San Roque and Bambi B Schieffelin. 2019. Perception verbs in context: Perspectives from Kaluli (Bosavi) child-caregiver interaction. *Laura Speed, C. O'Meara, Lila San Roque, and Asifa Majid (eds.) Perception Metaphors*, pages 347–368.

Frenette Southwood and Ann F. Russell. 2004. Comparison of Conversation, Freeplay, and Story Generation as Methods of Language Sample Elicitation. *Journal of Speech, Language and Hearing Research*, 47(2):366–376.

Eve Sweetser. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*, volume 54. Cambridge University Press.

John W. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Bram van Dijk, Tom Kouwenhoven, Marco Spruit, and Max Johannes van Duijn. 2023a. Large Language Models: The Need for Nuance in Current Debates and a Pragmatic Perspective on Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12641–12654, Singapore. Association for Computational Linguistics.

Bram van Dijk, Max van Duijn, Suzan Verberne, and Marco Spruit. 2023b. ChiSCor: A Corpus of Freely-Told Fantasy Stories by Dutch Children for Computational Linguistics and Cognitive Science. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 352–363, Singapore. Association for Computational Linguistics.

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.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing Pretrained Language Models for Lexical Semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

10. Language Resource References

van Dijk, Bram and van Duijn, Max and Verberne, Suzan and Spruit, Marco. 2023. *ChiSCor: A Corpus of Freely-Told Fantasy Stories by Dutch Children for Computational Linguistics and Cognitive Science*. Open Science Framework. PID https://osf.io/5h7za/?view_only=705c553e922046058ef9df52b4ac8ed7.

Representing Abstract Concepts with Images: An Investigation with Large Language Models

Ludovica Cerini[1], **Alessandro Bondielli**[1,2], **Alessandro Lenci**[1]
[1]Department of Philology, Literature and Linguistics, University of Pisa
[2]Department of Computer Science, University of Pisa
ludovica.cerini@phd.unipi.it
{alessandro.bondielli, alessandro.lenci}@unipi.it

Abstract

Multimodal metaphorical interpretation of abstract concepts has always been a debated problem in many research fields, including cognitive linguistics and NLP. With the dramatic improvements of Large Language Models (LLMs) and the increasing attention toward multimodal Vision-Language Models (VLMs), there has been pronounced attention on the conceptualization of abstracts. Nevertheless, a systematic scientific investigation is still lacking. This work introduces a framework designed to shed light on the indirect grounding mechanisms that anchor the meaning of abstract concepts to concrete situations (*e.g. ability - a person skating*), following the idea that abstracts acquire meaning from embodied and situated simulation. We assessed human and LLMs performances by a situation generation task. Moreover, we assess the figurative richness of images depicting concrete scenarios, via a text-to-image retrieval task performed on LAION-400M.

Keywords: LLMs, figurative language, multimodality, indirect grounding

1. Introduction

A naive view of abstract words regards them as expressing concepts that are not perceptually grounded, because they do not refer to entities with physical existence and perceivable by our senses. However, the relationship between abstract concepts and perceptual data is much more complex than it appears *prima facie*. Firstly, even if they lack direct grounding, abstract concepts are **indirectly grounded** (Louwerse, 2011; Dove, 2014; Utsumi, 2022), by being closely related to our embodied experience of the world (Paivio, 1990; Barsalou, 2008; Borghi et al., 2017). In this sense, abstract concepts succeed in acquiring perceptual representations via their association with concrete situations (e.g., the concept of *love* is grounded in the event of a mother hugging her child). Secondly, images can represent abstract concepts too. For instance, Figure 1 can be said to represent not only a mother with her baby, but also the concept of love. Actually, it is the association with abstract concepts that gives rise to metaphorical and figurative interpretations of images. Our hypothesis is that the ability of images to represent abstract concepts is determined by the indirect grounding of the latter: *An image represents an abstract concept A, if I depict a situation associated with A*. In this work, we use this hypothesis to address two main questions:

Q1 Do Large Language Models (LLMs) have human-analogue indirect grounding abilities to produce concrete situations that are strongly associated with abstract concepts?

Q2 Can the situations generated by the LLMs be used to retrieve images that represent abstract concepts? In addressing these questions, our main goal is to establish a framework aimed at enriching linguistic and multimodal resources for the study of metaphorical grounding of abstract concepts, with a focus on the Italian context. To this end, we selected a set of Italian abstract nouns and we set-up a **situation generation task** to compare the human-generated and LLM-generated situations produced in response to the target abstract word prompts. Then, we used the LLM-generated data in a **text-to-image retrieval task** from the LAION-400M dataset (Schuhmann et al., 2021a). Finally, the retrieved images were evaluated via crowdsourcing with respect to their ability to represent the target abstract concepts. The results of our experiments show that the ability of LLMs to ground abstract concepts on situations is very similar to the human one. Moreover, images retrieved through these situations strongly represent the target concepts used to generate them, suggesting that this method might be used to develop datasets of images annotated with their figurative meanings and to enhance the competencies of multimodal models to cope with metaphorical interpretations.[1]

2. Related Work

In linguistics, a number of approaches have been proposed to investigate the pragmatic abilities of

[1]We release the data collected across our experiments at https://github.com/lcerini/SituaMet (In preparation).

The Workshop on Cognitive Aspects of the Lexicon, pages 107–113
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Figure 1: Example of an image depicting a concrete situation evoking an abstract concept.

LLMs (Seals and Shalin, 2023; Hu et al., 2023). Barattieri di San Pietro et al. (2023) show that liguistic competence could be encoded distributionally in LLMs, thus allowing us to leverage such models to extract cognitive pattern of linguistic phenomena. This is particularly interesting for abstract conceptualization. Existing resources such as WordNet do not always reflect human representation of hierarchical relations among concepts (Bolognesi and Caselli, 2022; Liao et al., 2023). However, pragmatic inference involving meta-representional information still are not fully achieved in LLMs (Barattieri di San Pietro et al., 2023). Studies on non-literal understanding are still lacking, and the indirect grounding of abstract terms is understudied. Meta-representational, embodied, situated and multimodal aspects of conceptualization are the key to investigate metaphorical realization in a more complete and complex fashion. Major approaches from the NLP community in this context have been focused on text only (Shutova et al., 2010; Mohler et al., 2013; Shutova et al., 2016; Pramanick et al., 2018; Liu et al., 2020). Data scarcity and the cost of creating multimodal datasets impede research in this sense. Large datasets of image-text pairs, built by querying search engines, have been built to effectively train Large Vision-Language Models (VLMs) (Desai et al., 2021). However, due to their end goal, these dataset are less appropriate to study the more abstract and figurative aspects of images. Emerging approaches to multimodal figurative language have been proposed, and new effort has been made in the realization of multimodal metaphors datasets (Zhang et al., 2021; Akula et al., 2023). None, however, take fully into account the indirect grounding view of abstract concepts and images. Metaphorical images obtained by juxtaposition, resemblance, or fusion mechanisms instead lie outside the scope of the present work.

3. Situation generation task

The mental representation of abstract concepts makes use of associative relations to acquire meaning (Crutch and Warrington, 2005). To examine humans' abilities in grounding abstract concepts onto situations, and thus the capacity of LLMs to associate abstract concepts with real-world knowledge, we proceeded as follows. We first designed an elicitation task to investigate humans representation of abstract-related situations. Then, we translated it in a few-shot prompting task fed to a GPT-3 `davinci-003` model (Brown et al., 2020), which is the last non chat-oriented OpenAI GPT model and has shown state-of-the-art abilities in this regard.

Eliciting situations from humans To collect human data, we designed a test to elicit situations associated with 107 abstract stimuli. We divided the stimuli into three levels of concreteness (low, medium, and high) using norms from Brysbaert et al. (2013). *Low* abstract are concepts perceived as more anchored to concrete ideas or entities, while *high* abstract concepts are perceived as less anchored to concrete ideas or entities. *Victory* for example is seen as more concrete due to its association to experiences, unlike *justice*, which aligns with moral and social aspects.

Participants in a crowdsourcing experiment were asked to describe a situation that came to their mind given an abstract concept, with instructions including examples of situation formats to guide their responses. We submitted the test to 60 participants and obtained 539 situations in total, with an average of 5 situations per stimulus. We used Prolific[2] to crowdsource participants. An example of the abstract stimulus and the resulting situation is shown in Table 1.

Abstract Stimulus	Situation
Ability	Athletes performing acrobatic feats
Speed	A lion running

Table 1: Example situations generated by humans (translated from Italian).

Generating situations with an LLM To generate linguistic situations from abstract concepts with an LLM, we used the pre-trained `Davinci-003` GPT-3 model, following a structured output generation design. We exploited a few-shot prompting method, to obtain a specific format of generated situations, with $temperature = 0.5$ and $toppenalty = 1$. The few-shot prompt was constructed by using the same abstract stimuli used in the human elicitation task, followed by an arrow operator and two examples taken from human situations (See Fig. 2), ensuring consistency with the protocol used for human elicitation. We generated 10 situations per abstract concept, totalling 1070

[2] `prolific.co`

situations.

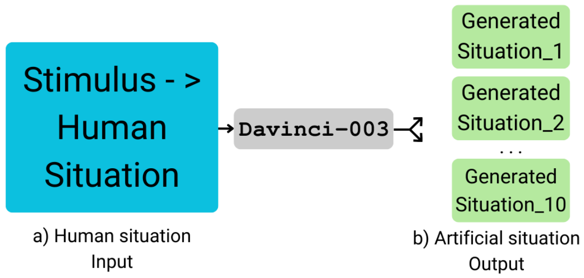


Figure 2: Few-shot prompt schema. Example of human input: *ability -> a person jumping*

Evaluation and Analysis Human-Elicited (HES) and Artificially-Generated (AGS) situations were then evaluated based on their similarity density and the associative strength to the abstract stimuli.

The qualitative analysis of HES revealed a certain degree of prototypical but diverse associations, i.e. similar situations among the different abstract concept groups. This is in line with literature indicating that people tend to anchor the representation of categories, simulating them in a typical perceptual situation (W.Yeh and Barsalou, 2006).

Starting from this, we explored how the two groups behaved in terms of typicality/diversity. We used `bert-base-italian-cased`, an Italian BERT model (Devlin et al., 2019), to obtain vector representations of the generated situations. For both HES and AGS groups we computed the the average cosine similarity between situations for each given abstract stimulus. A key aspect of the methodology involved constructing a dictionary that mapped each concept to its associated phrases. First, we iteratively processed the dataset and organized phrases according to their respective concepts. Then, we quantified the semantic similarity of phrases within each concept by computing the cosine similarity between each pair of phrases belonging to the concept. Note that we used the BERT tokenizer and right-padded all phrases to a standardized length (i.e., the maximum input size of the model), as it is customary when modelling texts with encoder-only Transformer architectures. We obtained embeddings for each phrase with a forward pass on the model, and then computed their pairwise similarity scores. Finally, we averaged the similarity scores to provide a quantitative measure of the semantic density of the phrases associated with the target concepts.

Figure 3 shows the average cosine similarities taking into account also the concreteness of the stimulus. Concreteness does not seem to affect similarity distributions. The median similarity values for the HES group was 0.365, while for the AGS group it was 0.375, indicating a slight central tendency difference between the two groups. As the two groups of items have been independently generated, we asses their statistical differences through a Wilcoxon test. The test showed a statistically significant difference between them ($W =$ 1203.5, $p <$0.05). However, the small difference in median values suggests that the magnitude of these differences might not be large in practical terms. The qualitative analysis of the results allowed us to identify several outliers that offer some insights into the nuances of human versus artificial sentence generation. In some concepts, AGS situations depict scenarios that, despite being distinct from each other, follow a similar structure in terms of the entities represented and the type of action involved. For example, the artificially generated sentences for the concept *power* predominantly feature situations involving entities associated with positions of power, such as *presidents, judges, and military figures*. This contrasts with the human-produced sentences, where more metaphoric situations also emerge, such as *a CEO's desk* versus an artificially generated scenario of *a CEO making decisions*. Similarly, in the case of the concept *peace*, human-generated sentences depict scenarios that do not necessarily involve animate entities (e.g., *a tranquil place*), offering a more abstract or symbolic representation of peace. Conversely, all artificially generated instances involve animate entities, such as *a group of people meditating together*. Despite the observed differences in similarity density and the varied entity types these outliers represent, they nonetheless appear to be associated with the abstract stimulus. To quantitatively evaluate this aspect and further understand how effectively HES and AGS capture and reflect the abstract concepts they are meant to represent, we proceeded to measure the associative strength. This measure aims to quantify the extent to which the generated phrases, regardless of their surface differences, retain a strong conceptual linkage to the original ideas they express.

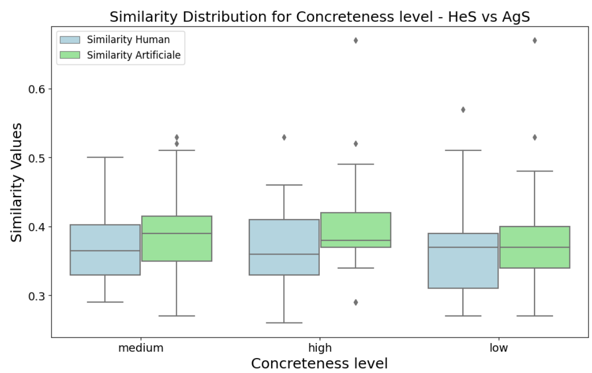


Figure 3: Similarities distribution in HES and AGS divided for concreteness level of abstract stimuli

To compute the associative strength between the stimuli abstract concepts the situations in HES and AGS we performed a rating assessment test using crowdsourcing via Prolific. Given an abstract concept and a situation obtained for it, we asked participants to rate how much the situation represents the paired abstract concept on a Likert scale from 1 (not at all) to 7 (very much). We collected ratings for all HES and AGS situations. Each concept- situation pair was rated by 10 participants, and we took the average rating. Figure 4 shows the distribution for each group. Both HES and AGS have a similar distribution, with median values 5.08 for HES and 5.07 for AGS ($\rho = 0.45$, p-value 0.00). This responds to **Q1** by suggestsing that i.) the LLM was able to generate human-analogue, coherent, and correct situations associated with the abstract stimuli, and that ii.) the generated situations could be considered as a proxy for the indirect grounding mechanism, providing evidence that diverse scenarios could suggest a figurative link between abstract concepts and real-world concrete events.

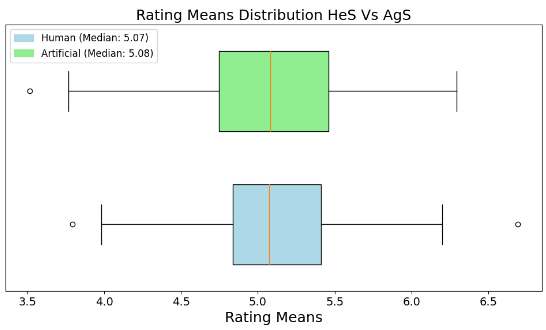


Figure 4: Rating avg. distribution in HES vs. AGS

4. Text-to-image retrieval task

Given the results of the previous experiment, we investigate on Q2 to understand whether i.) LLM-generated situations could be used to retrieve images able to represent an abstract concept and ii.) the same associative strength could be confirmed in the abstract-visual situation pair associations. To do so, we retrieved images based on the AGS, which were then evaluated via crowdsourcing.

Image Retrieval For our experiments we used the LAION-400M dataset (Schuhmann et al., 2021b). It contains 400 million of CLIP-filtered image-text pairs. We used AGS as queries for retrieving similar images from LAION-400M. To do so, we used CLIP (Radford et al., 2021) within the clip-retrieval library.[3]

[3]`https://github.com/rom1504/clip-retrieval`

For each situation in AGS, we collected the 10 most similar images. Thus, we select a total of 100 images per abstract concept and a grand total of 10,700 images. We set `aesthetic_score = 5` and `aesthetic_weight = 0.5` in clip-retrieval, to obtain images depicting real scenarios and limit the possibility of retrieving artistic illustrations.

Evaluation and Analysis To test whether querying LAION-400M with situations could retrieve images strongly associated with the target abstract concepts, we performed a rating assessment test. Further, we evaluated whether we can leverage the proposed retrieval methodology to analyze the connection between real-world scenarios and the figurative interpretation of images. We chose to test situations from AGS with an average rating above 5 (see Sec.3) to ensure that they were highly representative of the concepts for humans: 54 situations met this criterion. We selected the top-2 images per situation, for a total of 878 images. We employed two distinct multiple-choice tests for evaluation, gathering 60 participants via Prolific for each test (120 participants in total). The stimuli were divided into 12 sub-tests, with 5 annotators dedicated to each.

In *Test 1*, for each visual stimulus, participants were asked to label the image choosing between 4 options to label the image: the associated (i.e., correct) abstract concept and 3 other abstract concepts not associated with the image, used as distractors. In *Test 2*, we used different distractors. Specifically, we used a concrete word representing an object or entities depicted in the image, and an abstract word and a concrete word not associated with the image. Distractors were chosen to exclude synonyms or concepts directly related to the correct option. This format allows us to assess the strength of the relationship between the abstract concept and the image, while still presenting options that could be plausibly related. We adopetd a multiple-choice framework that incorporates distractor options for its robustness against possible bias and vagueness of answers. In fact, visuals inherently carry diverse interpretations, enabling a singular image to be associated with numerous abstract concepts. Introducing a variety of abstract choices allows for a more precise evaluation of the strength of the association between the correct abstract stimulus and the image. Using a rating system that includes only the correct notions could potentially introduce bias, as it pre-defines the links between images and abstract concepts. On the contrary, an open-ended response format might elicit a wide array of answers, which could be less beneficial in analyzing the connection between specific abstract-visual scenario pairs. The lack of constraints in responses to tasks centered on abstract concepts would have resulted in extremely varied outcomes. This is due to the complex nature of abstract concepts and their associative paths, which may manifest differently across contexts and intensities. Typically, unrestricted tasks lead participants to associate abstract concepts with synonyms or to describe images in literal terms initially. Given the study's goal to delve into a specific representation of abstract concepts, specifically their metaphorical interpretation through images, we believe that the use of distractors serves to probe these kinds of associative links more effectively.

Table 2 provides the results for both tests. For Test 1, when we look at *labels distribution*, we see that participants chose the correct abstract concept as label in most cases (69,3%), suggesting a degree of association between the two. To further evaluate their *association strength*, we proceeded as follows: we binarized the results based on whether the associated abstract concept was chosen by at least 50% of participants across all images for the same stimulus. In this case, we can suppose that the images retrieved could be interpreted figuratively via indirect grounding. This evaluation shows that 90.7% of images are vehicle of the associated abstract concepts. In test 2, results indicate that, notwithstanding the presence of a concrete term corresponding to the visual representation in the image, participants selected the correct abstract labels in 29.1% of instances. In this 29.1%, 5.6% have more than 50% of correct abstract labels. These results confirm the variability of visual semantics. Nevertheless, still a fair enough percentage of correct abstract labels were assigned. We can argue that our findings suggest a positive answer for Q2.

	Test 1		Test 2	
	Corr.	Incorr.	Corr.	Incorr.
Labels distribution	**69.3%**	30.7%	29.1%	70.6%
Association strength	**90.7%**	9.3%	5.6%	94.4%

Table 2: % of correct/incorrect labels.

5. Discussion and Conclusions

In this work, our primary aim was to provide a framework for building or enriching linguistic/multimodal resources to delve into the metaphorical grounding of abstract concepts, focusing on its applicability within the Italian linguistic context. We evaluated the abilities of humans and a LLM, namely `Davinci-003` GPT-3 model to generate situations that could provide indirect grounding to abstract concepts.

Our first experiment suggests that LLMs-generated situations are comparable to those generated by humans. Similarity density analysis suggests that the AGS align with the cognitive system's ability to produce diverse conceptualizations for an abstract concept (Barsalou, 2003). The proposed text-to-image retrieval method confirmed that images depicting situations grounding abstract concepts can represent these concepts, adding a small but significant piece to the indirect grounding theories that still lack empirical evidence (Utsumi, 2022). However, the evaluation task conducted with concrete distractors, confirms the idea that visual properties of an image do not always coincide with linguistic properties (Giunchiglia et al., 2023), generating a mismatch between abstract and concrete classification. By analyzing the images that struggled to recall the associated abstract concepts, we also found that retrieved images for these situations (i.e., the AGS query) is hard to be visually represented. For example, the AGS *a child having to face the loss of a parent* generated for the concept *adversity* has retrieved the image of a sad girl, which only partially represents the scenario depicted in the generated sentence. The proposed method may prove to be beneficial for at least two reasons. First, its direct outcome is a way to automatically obtain metaphorically-rich images from dataset aimed at Computer Vision or Language-Vision problems, by exploiting LLMs generative abilities and VLM-based retrieval. Second, this kind of data may become a valuable asset in facing the lack of metaphorical multimodal dataset, to achieve a better understanding of the indirect grounding mechanisms in a multimodal setting. Moreover, it could enrich the abstract concepts understanding capabilities of VLMs by training on metaphorical- and abstract-oriented data. To these ends, we release all the data collected for the present work. Our future plans include expanding the present work by further exploring prompting techniques and the use of LLMs and VLMs. Moreover, we intend to adopt the proposed method to augment existing datasets, either linguistic, visual and multi-modal, with information concerning abstract concepts and figurative interpretations. To enhance our framework, we also plan to broaden the set of abstract concepts beyond the initial 107, and to incorporate more comprehensive measures beyond similarity density for evaluating the generated data. Additionally, we aim to test the framework across various linguistic and cultural systems. We believe that the proposed research framework can indeed be utilized to explore the phenomenon of abstract concept representation via images in other languages as well. This would allow us to ascertain whether cross-cultural differences emerge in the anchoring of abstract terms to situations, or whether similar patterns of metaphorical grounding are observed across different linguistic and cultural landscapes.

Acknowledgments

This research was partly funded by PNRR—M4C2—Investimento 1.3, Partenariato Esteso PE00000013—"FAIR—Future Artificial Intelligence Research"—Spoke 1 "Human-centered AI," funded by the European Commission under the NextGeneration EU programme.

Bibliographical References

Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. 2023. Metaclue: Towards comprehensive visual metaphors research.

Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The pragmatic profile of chatgpt: Assessing the communicative skills of a conversational agent. *Sistemi intelligenti, Rivista quadrimestrale di scienze cognitive e di intelligenza artificiale*, (2/2023):379–400.

Lawrence Barsalou. 2003. Situated simulation in the human conceptual system. *Language and Cognitive Processes*, 18(5-6):513–562.

Lawrence Barsalou. 2008. Grounded cognition. *Annual review of psychology*, 59:617–45.

Marianna Bolognesi and Tommaso Caselli. 2022. Specificity ratings for italian data. *Behavior Research Methods*, pages 1–18.

Anna Borghi, Ferdinand Binkofski, Cristiano Castelfranchi, Felice Cimatti, Claudia Scorolli, and Luca Tummolini. 2017. The challenge of abstract concepts. *Psychological Bulletin*, 143.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marc Brysbaert, Amy Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46.

Sebastian Crutch and Elizabeth Warrington. 2005. Abstract and concrete concepts have structurally different representational frameworks. *Brain : a journal of neurology*, 128:615–27.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: web-curated image-text data created by the people, for the people. *CoRR*, abs/2111.11431.

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.

Guy Dove. 2014. Thinking in words: Language as an embodied medium of thought. *Topics in cognitive science*, 6.

Fausto Giunchiglia, Mayukh Bagchi, and Xiaolei Diao. 2023. Aligning visual and lexical semantics. In *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity*, pages 294–302, Cham. Springer Nature Switzerland.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models.

Jiayi Liao, Xu Chen, and Lun Du. 2023. Concept understanding in large language models: An empirical study.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 250–255, Online. Association for Computational Linguistics.

Max M. Louwerse. 2011. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302.

M. Mohler, D. Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. pages 27–35.

A. Paivio. 1990. *Dual Coding Theory. In Mental Representations: A Dual Coding Approach*. Oxford: Oxford University Press.

Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An LSTM-CRF based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67–75, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021a. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021b. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

S. M. Seals and Valerie L. Shalin. 2023. Long-form analogies generated by chatgpt lack human-like psycholinguistic properties.

Ekaterina Shutova, Lin Sun, Elkin Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. 2016. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43:1–88.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. volume 2, pages 1002–1010.

Akira Utsumi. 2022. A test of indirect grounding of abstract concepts using multimodal distributional semantics. *Frontiers in psychology, 13, 906181*.

W.Yeh and LW. Barsalou. 2006. The situated nature of concepts. *J Psychol.*, pages 119(3):349–84.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, Online. Association for Computational Linguistics.

Big-Five Backstage: A Dramatic Dataset for Characters Personality Traits & Gender Analysis

Marina Tiuleneva[1*], Vadim Porvatov[2], Carlo Strapparava[1]
[1]University of Trento, [2]University of Amsterdam
[1]Via Calepina 14, 38122 Trento, Italy,
[2]Science Park 904, 1098 XH Amsterdam, The Netherlands
[*]mari.tyuleneva@gmail.com

Abstract

This paper introduces a novel textual dataset comprising fictional characters' lines with annotations based on their gender and Big-Five personality traits. Using psycholinguistic findings, we compared texts attributed to fictional characters and real people with respect to their genders and personality traits. Our results indicate that imagined personae mirror most of the language categories observed in real people while demonstrating them in a more expressive manner.

Keywords: psycholinguistics, dataset, natural language processing, Big-Five personality traits, LIWC, MRC

1. Introduction

Can fictional characters be written so skillfully as to be indistinguishable from real people? Reading fiction opens up the inner worlds of the characters, their experiences and emotions, allowing the reader to take part in their life lessons, enhancing imagination and social competence (Boyd, 2009). It has been experimentally shown that reading different types of literary genres influences the social cognition of the reader (Kidd and Castano, 2013), (Heyes, 2018). One of the creative aspects of fiction that enables readers to immerse themselves in a character's perspective is the unique ability of the author to compose dialogue resonating with the authenticity of human speech. Previous research analysing a limited number of theatre plays written in verse (Ireland and Pennebaker, 2010) and movie scripts (Nalabandian and Ireland, 2022) has shown that certain authors can successfully imitate real people's speech, while others intentionally or not fail to do so (e.g., Shakespear's female characters speak like men; the same thing can be seen in Woody Allen movies).

The exploration of how fiction mirrors real-life speech finds its foundation in the distinct traits that individuals exhibit in their communication, both spoken and written. A gender-specific vocabulary-based approach has shown persistent differences in language use by males and females (Pennebaker and King, 1999). These findings were further confirmed on a big corpus of various types of texts (Newman et al., 2008): for example, women tend to use more emotional words and negations than men, and express thoughts, emotions and senses to other people. In contrast, men typically refer to external occurrences, objects, and processes, as well as utilize technical linguistic elements (numbers, articles, prepositions, and long words).

Further, the linguistic properties of texts written by people with different personalities have been extensively studied for the Big-Five taxonomy (Mairesse and Walker, 2007). This framework is centered around 5 major personality traits (Goldberg, 1990): *Extraversion* (EXT), *Neuroticism* (NEU), *Agreeableness* (AGR), *Conscientiousness* (CON), and *Openness* (OPN). For example, results show that informal speech is more common for extraverts than for introverts (*Hi* vs. *Hello*), neurotics more often use negative emotional vocabulary and first-person singular pronouns (*I*, *my*), conscientious individuals avoid negations, open people prefer longer words and vocabulary related to curiosity. These findings have also been confirmed for texts written on social media (Mewa, 2020).

Such discoveries from psycholinguistic research can be applied as a starting point for comparative analysis of authentic texts produced by real people and text written for fictional characters (Picca and Pitteloud, 2023). Studying imagined personae brings insights into the properties of separate works produced by the same author, whose intent is to mimic natural communication (Boyd and Pennebaker, 2015). Knowing patterns in the behaviour of fictional characters can give us a better understanding of sociocultural norms, and the extent to which it is possible for professional writers to imitate real-world speech.

The primary goal of this research is to evaluate the capability of authors to produce texts that convincingly mimic the speech of different genders and personalities. We focus on theater plays by internationally renowned authors as our primary source, as these plays rely on direct speech for character portrayal. Previous studies have mostly focused on movie scripts and have not investigated other narratives, such as those belonging to literary fiction.

The Workshop on Cognitive Aspects of the Lexicon, pages 114–119
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

	EXT	AGR	OPN	NEU	CON
Acc.	0.872	0.889	0.894	0.791	0.9
Pr.	0.67	0.949	0.807	0.95	0.838
Rec.	0.88	0.862	0.852	0.663	0.94
F1	0.76	0.903	0.829	0.781	0.886

Table 1: GPT-3.5 performance measured by accuracy, precision, recall, and F1-score

		EXT	AGR	OPN	NEU	CON
M	0	0.4	0.26	0.44	0.37	0.35
	1	0.21	0.35	0.17	0.24	0.25
F	0	0.28	0.11	0.27	0.19	0.26
	1	0.11	0.28	0.12	0.2	0.13

Table 2: Big-Five personality traits distribution between male and female characters

In order to further boost research in this area and extend it to new domains, we prepared the Big-Five Backstage dataset comprised of fictional characters lines. To demonstrate its potential, we performed character analysis based on their genders and Big-Five personality traits. Character comparisons based on linguistic categories have shown that fictional males and females generally repeat language patterns observed in real people. The same trend can be seen for personality traits. Moreover, we found that specific language categories demonstrate a more drastic difference in imagined personae than in real people.

2. Data

2.1. Data Extraction & Preprocessing

The raw data consisted of 178 files containing theatre plays downloaded from the Project Gutenberg website. After having excluded non-English literary works along with those composed in verse, 400 theatre plays remained, written by 132 different authors. Next, we extracted the lines belonging to each character in the plays, and excluded the ones with fewer than 5 lines. The obtained text was normalized and tokenized with the help of the Stanza framework (Qi et al., 2020). The resulting dataset consists of 3 265 text samples corresponding to the concatenation of lines spoken by each character. Overall, it contains 3 419 136 words with a mean equal to 1047.2 words per character. The auxiliary part of the dataset includes author-level labels reflecting their gender, country of origin, and years of life.

2.2. Annotation Process

Each character was manually labelled as *male* (M) or *female* (F). For Big-Five personality traits annotation, GPT-3.5 (gpt-3.5-turbo) was applied to label each trait. These results were further compared with human annotations in order to validate predictions. The choice of GPT as an annotation tool was informed by prior research indicating the capacity of Large Language Models (LLM) to properly mimic the language of individuals across different personalities and genders, showing linguistic characteristics aligned with real human speech (Jiang et al., 2023). To achieve this, it is sufficient to create a prompt that clearly specifies certain personality traits and gender, according to which the LLM can generate text that reflects the linguistic patterns of specified traits according to the Five Factor Model (McCrae and John, 1992).

We sampled 10% of texts from the initial dataset and randomly distributed them among two human annotators. These annotators were blind to the GPT-3.5 responses and were subsequently instructed to read the texts and categorize them based on the Big-5 personality traits, marking a '1' if a trait was perceived in the text, and '0' if it was not. Definitions of the Big-5 personality traits were adopted from (Mairesse and Walker, 2007) (e.g., "Extraverts are sociable, assertive, playful, whereas Introversion is characterized by being aloof, reserved, shy."). Consequently, each text received annotations from one human and the GPT-3.5. We provide the following example that shows GPT-3.5 was not always able to correctly understand the intent of the characters. Here, in text id=94 GPT-3.5 misidentifies the Extraversion trait and labels the character as an introvert:

> **Text:** <...> the fact is, Dick, I'm lonely in this great house. You, who pass most of your time in a store, with people constantly coming and going, take pleasure in spending a week or two in a quiet place. <...>
>
> **GPT-3.5**: *Extraversion 0*. The person mentions feeling lonely in their house, indicating a lack of social interaction and possibly a preference for solitude.
>
> **Ground truth**: *Extraversion 1*. Experiencing loneliness does not imply a preference for solitude.

To evaluate the preliminary annotation's effectiveness, we assessed the performance of GPT-3.5 in terms of accuracy, precision, recall, and F1-score, as presented in Table 1. Although accuracy varied across traits, the lowest F1-scores were recorded for *Extraversion* and *Neuroticism*. Table 2 displays the final distribution of traits among the characters after validation by humans.

Additional information about the GPT-3.5 ablation study is available at the corresponding part of our GitHub repo.

3. Methods

We employ psychological findings that have identified connections between a range of linguistic markers (LMs) and genders (Pennebaker and King, 1999), as well as LMs and personality traits (Mehl et al., 2006). Linguistic markers represent clusters of words with a common characteristic, such as pronouns (*i*, *you*, *that*), prepositions (*to*, *of*, *in*), social (*you*, *we*, *he*, *she*) and cognitive (*but*, *know*) processes.

In order to study differences in texts of fictional men and women, we use the Linguistic Inquiry and Word Count (LIWC) dictionary (Boyd et al., 2022) and choose 44 LMs proposed in (Newman et al., 2008). This work analyses various groups of texts, finds LMs showing statistically significant differences between male and female writings, and presents Cohen's *d* coefficient (Cohen, 1992) obtained for each category of words. We compute frequencies of the LMs and compare Cohen's *d* of those showing statistical significance with Cohen's *d* calculated for real people. This analysis was performed for the whole dataset and for individual authors with at least 10 characters of each gender. On top of that, we extend the linguistic comparison to the characters' personality traits by applying MRC Psycholinguistic Database markers proposed in (Mairesse and Walker, 2007) in the addition to the mentioned subset of LIWC.

We took the results of statistical tests performed on real people texts from (Mairesse and Walker, 2007) and (Newman et al., 2008). In (Mairesse and Walker, 2007), the authors calculated Pearson's correlation coefficients between LIWC/MRC features and personality traits, while (Newman et al., 2008) provides the word frequencies and the effect size for the most common groups of words used by men and women.

During the analysis of the provided texts, we apply several methods from classic statistics: Mann-Whitney U test (Mann and Whitney, 1947) and Wilcoxon signed-rank test (Wilcoxon, 1945) for sample difference testing, Cohen's *d* coefficient to quantify the discovered differences, and point-biserial correlation (Lev, 1949) as a measure of dependency between LMs' frequencies and Big-Five personality traits considered as dichotomies. For all of the applied tests, we consider the level of significance α = 0.05.

4. Results

4.1. Genders

We performed the Mann-Whitney U test for male and female populations across the dataset and found 32 LMs that show statistical significance. Next, we calculated Cohen's *d* for these markers and compared them to real people. The difference in LMs between fictional characters repeats the patterns in the real world: men show a preference for long words (*BigWords*, >6 letters; d=0.33), prepositions (d=0.29), work-related vocabulary (d=0.23), numbers (d=0.2) and swear words (d=0.13), while women utilize language categories related to family (d=−0.36), home (d=−0.2) and social processes (d=−0.19), use pronouns (d=−0.25) and negations (d=−0.2). We also found LMs that show an opposite trend to the real world, which we call *reversed markers*. One such category, the second-person pronoun *you* (d=−0.1), occurs more often in fictional female speech, whereas in the real word it is used more often by men.

As shown in previous examples, the effect size for a number of LM's meets Cohen's *d* criteria for small $(0.1 \leq |\mathrm{d}| < 0.3)$ and medium effect $(0.3 \leq |\mathrm{d}| < 0.5)$, and for all of them it exceeds Cohen's *d* for real people, Figure 1(a). This indicates that there is an exaggerated difference for both men and women in fiction. Therefore, we continued our research focusing on individual authors, excluding those having fewer than 2 characters of each gender. We show the top-10 authors whose usage of LMs follows the same patterns as has been reported for real people, Figure 2.

For all the authors with at least one statistically significant LM, we calculated Cohen's *d* and did another comparison to real people. Thus, we confirmed the presence of an author-level exaggeration of gender-specific markers for males and females. In order to measure this effect, one can utilize the coefficients of a linear regression based on Cohen's *d* values for LMs, as shown for characters of August Strindberg, Figure 1(b). The slope of the linear regression line indicates the level of hyperbolization for both genders while the intercept sign demonstrates an imbalance in favor of females (negative) or males (positive). Conducted measurements on the sample of authors allow us to report that the mean value for the slopes is 4.5 with Q1=2.5 and Q3=5.5 while the mean value for the intercepts is −0.169 with Q1=−0.33 and Q3=−0.024. This indicates that the exaggeration is pronounced and slightly disproportional towards women.

4.2. Big-Five Personality Traits

We use 65 LMs (51 from LIWC, 14 from MRC) to analyze the linguistic differences in the personality

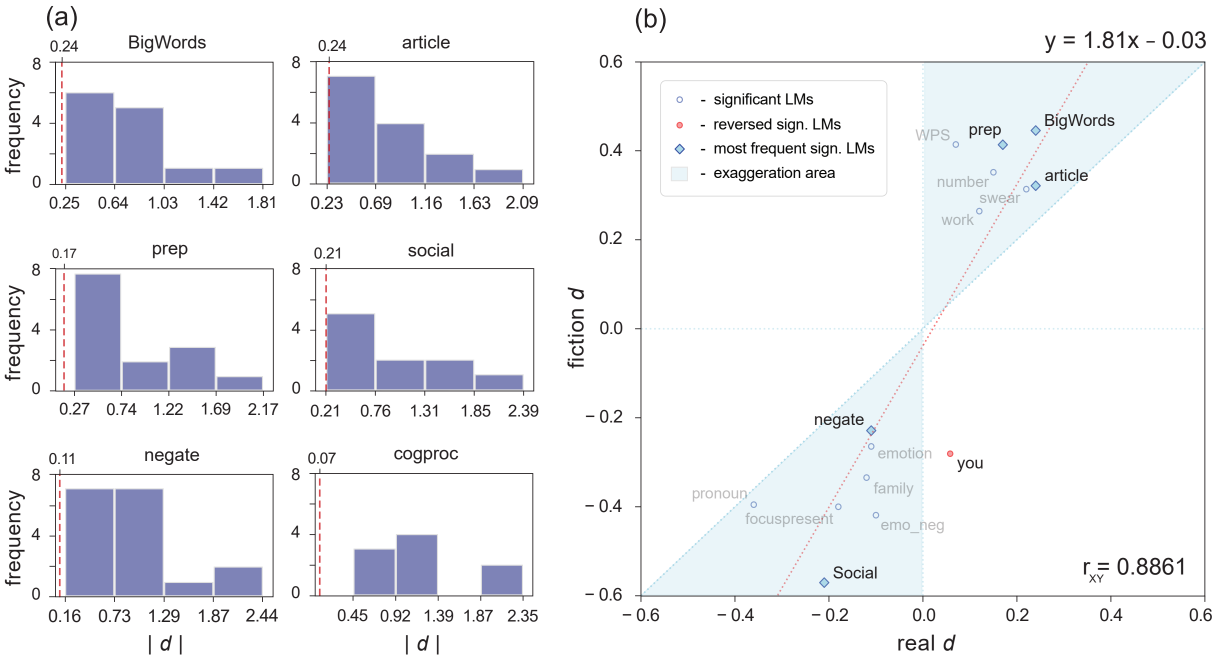


Figure 1: (a) Effect size distribution in 6 most frequent linguistic markers for the whole dataset: the red line shows Cohen's *d* for a corresponding LM in real people. (b) Example of author-level correlation between Cohen's *d* calculated for statistically significant LMs in real people and fictional characters. As Cohen's *d* is based on a mean difference of two samples, its positive values show that males used certain LM more than women, while the negative ones suggest the contrary.

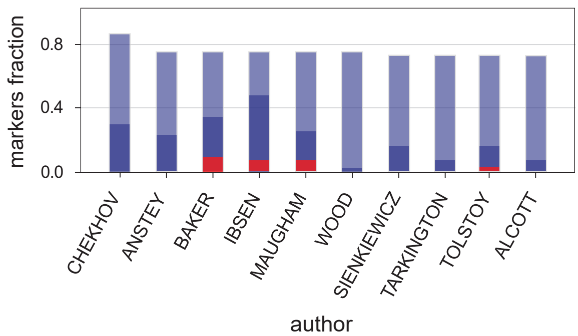


Figure 2: Fraction of linguistic markers indicating the difference between males and females (per author/ total LMs): real-world pattern of use (light blue), statistically significant (dark blue), reversed (red).

traits of characters. A significant point-biserial correlation was observed at least in one trait for all 14 MRC and 48 LIWC linguistic markers. We found 5 LMs showing statistical significance in all of the traits, Figure 3(a), and 24 LMs that are significant for 4 traits, Figure 3(b).

The strongest positive correlation is found for *Neuroticism* and LIWC markers corresponding to word count (*WC*) and linguistic categories related to affective vocabulary (*emo_neg*, *emo_anger*, *death*). In contrast, emotionally stable characters show a preference for punctuation marks (*AllPunc*). The pronounced dependency for *Conscientiousness* as a trait showing self-discipline was found in the case of the MRC summary variables: number of letters in one word (*NLET*) and number of phonemes (*NPHON*). Otherwise, unconscientious personae are typically depicted in plays by using excessive punctuation, such as exclamations, quotation marks, and non-fluent words (*nonflu*: *oh*, *um*). Similarly to real people, fictional extraverts communicate through vocabulary related to leisure, whereas introverts show a preference for long sentences. Agreeable characters tend to use positive emotional words (*emo_pos*), and their opposites rely on the negative ones (*emo_neg*, *emo_anger*). Finally, the presence of *Openness* correlates with Paivio's Meaningfulness (*MEANP*), spelling (*NLET*), and leisure. It also has the most discrepancies with the texts attributed to real people due to the largest number of reversed markers among the traits.

5. Limitations

This study employs GPT-3.5 as a tool for annotating Big-Five personality traits in textual data, complemented by the analysis of a single human annotator. A limitation of our methodology arises from the uncertainty surrounding the actual personality traits of the texts belonging to the fiction characters under examination. The ground truth cannot be established due to the nature of this data, and our

Figure 3: LMs demonstrating statistically significant correlation for 5 traits (a), and for 4 traits (b). The intersection of LM and trait represents a value of biserial point correlation between them; linguistic markers are sorted by the mean of correlations across all traits with asterisks denoting the reversed markers. The bar chart shows the fraction of authors that use a certain LM in at least one trait.

analysis operates under assumptions made by both the AI and human annotator based on the textual evidence available. Furthermore, the reliance on a single human annotator introduces potential biases and a lack of diverse interpretative perspectives that multiple annotators could provide. This limitation could affect the reliability and generalizability of the findings, as the interpretation of personality traits from text is subjective and may vary significantly among different readers.

6. Discussion

Our study presents a methodological framework that offers valuable insights not only into theater plays but also extends to real-life contexts. This framework has potential applications in various domains such as social media (for authorship attribution and detecting anomalous behavior) and cultural studies (exploring gender and social stereotypes, and analyzing authors through their characters). In the area of Human-Computer Interaction and robotics, our dataset and methodology could prove instrumental in assessing texts generated by large language models.

We have provided a statistical analysis of word usage shifts in theatrical texts across previously unexplored dimensions. While related research has focused on verse-based theater plays by a select group of authors (Ireland and Pennebaker, 2010), our study pioneers in examining the extent to which authors can replicate the speech of male and female characters and differentiate their characters from real individuals. Utilizing LIWC/MRC dictionaries, we observed that specific word categories correlate with certain personality traits, in line with prior studies (Mairesse and Walker, 2007), (Kosinski et al., 2013). Interestingly, our findings highlight a tendency among authors to overemphasize gender-specific vocabulary, particularly in depicting female speech. This suggests that while some authors successfully mirror real-world linguistic trends, others struggle to accurately represent these nuances in their characters.

We have also identified correlations between linguistic markers and personality traits, revealing dependencies for further investigation. For instance, emotive vocabulary is linked with Neuroticism, Extraversion, and Agreeableness, while summary variables can distinguish Conscientiousness, and specific punctuation usage is common among unconscientious and emotionally stable personalities.

Our findings underscore the challenge authors face in naturally replicating real speech patterns. Even when attempting to 'mimic' individuals of different genders, authors often exaggerate certain speech characteristics. Characters portraying various personality types exhibit more pronounced linguistic features than typically observed in real individuals. Our research invites further exploration into the nuances of generating speech that aims to mimic another's, whether by humans or machines.

In conclusion, our study demonstrates the potential of an automated approach for labeling Big-Five traits. Moving forward, we aim to delve deeper into the zero-shot capabilities of large language models in predicting personality traits, highlighting the need for more research in this area to refine and expand upon our promising results.

Acknowledgements

We thank Fredrik Hoeglund for help with the annotation process.

7. Bibliographical References

Brian Boyd. 2009. *On the origin of stories: Evolution, cognition, and fiction*. Harvard University Press.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Ryan L Boyd and James W Pennebaker. 2015. Did shakespeare write double falsehood? identifying individuals by creating psychological signatures with text analysis. *Psychological science*, 26(5):570–582.

Jacob Cohen. 1992. Statistical power analysis. *Current directions in psychological science*, 1(3):98–101.

Lewis R Goldberg. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Cecilia Heyes. 2018. *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.

David Comer Kidd and Emanuele Castano. 2013. Reading literary fiction improves theory of mind. *Science*, 342(6156):377–380.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805.

Joseph Lev. 1949. The point biserial coefficient of correlation. *The Annals of Mathematical Statistics*, 20(1):125–126.

François Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. pages 496–503.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862.

Tasneem Mewa. 2020. 'personality, gender, and age in the language of social media: The open-vocabulary approach'by h. andrew schwartz et al (2013). *Identifying Gender and Sexuality of Data Subjects*.

Taleen Nalabandian and Molly E Ireland. 2022. Linguistic gender congruity differentially correlates with film and novel ratings by critics and audiences. *PloS one*, 17(4):e0248402.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse processes*, 45(3):211–236.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Davide Picca and Jocelin Pitteloud. 2023. Personality recognition in digital humanities: A review of computational approaches in the humanities. *Digital Scholarship in the Humanities*, page fqad047.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

8. Language Resource References

Qi, Peng and Zhang, Yuhao and Zhang, Yuhui and Bolton, Jason and Manning, Christopher D. 2020. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. The Stanford NLP Group. PID https://github.com/stanfordnlp/stanza.

Interaction of Semantics and Morphology in Russian Word Vectors

Yulia Zinova, Anastasia Yablokova, Ruben van de Vijver
Heinrich-Heine-Universität
Universitätsstraße 1, 40225 Düsseldorf
zinova@hhu.de, anyab100@uni-duesseldorf.de, ruben.vijver@hhu.de

Abstract

In this paper we explore how morphological information can be extracted from fastText embeddings for Russian nouns. We investigate the negative effects of syncretism and propose ways of modifying the vectors that can help to find better representations for morphological functions and thus for out of vocabulary words. In particular, we look at the effect of analysing shift vectors instead of original vectors, discuss various possibilities of finding base forms to create shift vectors, and show that using only the high frequency data is beneficial when looking for structure with respect to the morphosyntactic functions in the embeddings.

Keywords: Russian, FastText, Embeddings, Morphology, Semantic Classes, Syncretism

1. Background

Learners of morphology, whether humans or machines, must be able to overcome the Zipfian distribution of words: A few words occur extremely frequently, and most words are very infrequent (Kodner, 2022; Guzmán, 2020). As a result, there are many words that a learner has to produce, although these words have never been encountered in the input. A question that arises is what allows a learner to do this (Guzmán, 2020; Ackerman and Malouf, 2013)?

Some proposals to achieve this have focused on the form-side of morphology (Ackerman and Malouf, 2013; Albright, 2010; Malouf, 2017). These proposals leverage implications among forms in a paradigm. For example, the Latin genitive form of King *regis* allows a learner to predict most other forms of the paradigm (the dative form *regi*, the accusative *regem* and the ablative *rege*), whereas the nominative form *rex* does not. So, if the learner knows the genitive form, they can use this knowledge to predict all other forms (see Albright, 2010, for an explanation of how this mechanism affected diachronic changes in Yiddish).

Yet, it is not clear whether language users really use forms to produce other forms (Nieder et al., 2021a,b,c). One reason is that in most languages it is not obvious what is the most informative form. Finnish, for example, has several forms that can be used to base other forms upon (Nikolaev et al., 2022b). Moreover, this focus on form alone neglects any role of semantics in predicting the meaning of words that have not been encountered.

Information about the semantics of words can be captured by embeddings, and can be used to investigate properties of paradigms that can be helpful to learners. Recent work on morphology used the information contained in embeddings to investigate specific properties of morphology. For example, Westbury and Hollis (2019) has used embeddings to investigate whether part-of-speech can be predicted from embeddings.

Embeddings are learned in an unsupervised manner from raw text and thus contain information about the distribution of words in a corpus (Bojanowski et al., 2017; Landauer and Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014). An additional difficulty for languages with rich morphology, such as Russian, is the huge amount of out-of-vocabulary words. This problem is addressed by fastText word vectors (Bojanowski et al., 2017), since the model learns representations of the n-grams and thus a representation of any given word is either directly learned or calculated on the basis of the n-gram representation of its parts.

FastText representations have been shown to work well for many tasks, however, there is potential for improvement. The idea of representing the word through the sum of its n-grams relies on the idea that affix n-grams correspond to functions and these functions can be represented in a similar way as the whole words. It has been shown by Nikolaev et al. (2022a) and Shafaei-Bajestan et al. (2022) that though this idealisation works, it is not accurate to represent a single morphosyntactic function with a unique vector: there is interaction between various functions (Nikolaev et al., 2022a) as well as between a function and semantic classes (Shafaei-Bajestan et al., 2022). An additional problem in Russian is syncretism: same affixes can be encountered in various cells within a paradigm as well as represent different functions across the paradigms. For example, the genitive and accusative of the word for an animate masculine noun ‘elephant’ are both *slona* and at the same time for example the nominative singular of an animate feminine noun ‘mother’ *mama* have the same affix. As a result, the embedding of the word *slona* necessarily contains distributional information about its occurrence in genitive and in accusative contexts and the representation of a trigram ending on *a* at the end

The Workshop on Cognitive Aspects of the Lexicon, pages 120–128
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

of the word may, in addition, contain distributional information about nominative singular contexts.

In order to generate reliable predictions for the meaning of each morphosyntactic function of syncretic forms, we propose to extend a method proposed in (Nikolaev et al., 2022a; Shafaei-Bajestan et al., 2022). Instead of assuming that the meaning of a word is best predicted by the sum of the vectors of its n-gram components, we represent as a sum of the embeddings for the base and the morphosyntactic functions it expresses. In order to do so, we use dictionaries and existing morphological tools in order to find out-of-vocabulary word forms. We also use the word forms for which we have embeddings to extract vectors for lemmas and grammatical functions. We then use the inferred vectors to predict vectors for out-of-vocabulary word forms, to provide multiple vectors for syncretic forms. At the current stage we rely on dictionary information in order to explore such representation, but our final aim is a self-supervised pipeline.

According to Wiemerslage et al. (2022) the next challenge in computational morphology is to understand morphology from text alone. Wiemerslage et al. 2022 introduces the task of *truly unsupervised morphological paradigm completion* and proposes a pipeline for approaching it. In a first step, Wiemerslage et al. (2022) clusters word forms into paradigms on the basis of their orthographical similarity. In a second step, it is assessed which orthographic changes on the word forms express the same inflectional information. For example the last character in the Russian word *okna* 'windows' and the last one in *mamy* 'mothers' express the same inflectional information (namely, nominative plural). Information about word embeddings is then used to assess the distribution of such inflections, and this, in turn, is used to assign labels to word forms. These labeled word forms are then used to train a morphological learner. The model presented in Wiemerslage et al. 2022 is trained on digitized children's books and the Bible in several languages (German, Greek, Icelandic, and Russian). The evaluation has been done in terms of correct paradigm reconstructions with paradigm slots aligned between different lemmas but in random order, the best possible correspondence to true labels being selected for the evaluation. The best results across all the languages and training data are about 27% correctly generated word forms for Russian digitized children's books. The pipeline proposed in Wiemerslage et al. 2022 is in principle unable to cope with syncretism, since any string can be mapped to only one functional slot. This raises the question how morphology can be learned in an self-supervised way while also taking into account the fact that a lot of languages exhibit syncretism.

While the pipeline of the Wiemerslage et al. (2022) works with the original vectors, comparisons among vectors yields further vectors, and there have been proposals in the literature to look at the structure of such comparisons instead (Nikolaev et al. 2022a, Shafaei-Bajestan et al. 2022). For example, one could assume that the vector of one word form, which we refer to as the base vector, in a paradigm is used to derive other word forms. An obvious choice for the base vector would be the nominative singular form, since it is the base form provided in the dictionaries. But the nominative case often is syncretic in Russian, which becomes especially concerning when working with other forms, with which the nominative singular is syncretic. Furthermore, the same dictionaries often list a set of other forms (principal parts) to provide the full information needed to reconstruct the paradigm, which may serve as an indicator that nominative singular alone may not be enough for our purposes. In the following we will investigate various choices for a base vector.

2. Methodology

2.1. Data

To explore the semantic space of Russian nouns, we first need an overview of the nominal paradigms. We obtained it by extracting 14,157 nouns from a recent frequency dictionary (Ljaševskaja and Šarov, 2009) together with their frequency information. These nouns were parsed using the `pymorphy2` library (Korobov, 2015) and inflected along the list of fourteen forms: seven cases and two numbers. The cases include the six standard cases as well as the second genitive (partitive), here with the abbreviations used further throughout the paper: *nominative (nom), genitive (gen), dative (dat), accusative (acc), ablative (abl), locative (loc), genitive 2 (gen2)*. Each of these cases occurs in the *singular (sg)* or *plural (pl)*, as in Table 1.

We have excluded nouns for which *pymorphy* could not find a parse as well as those where not all of the paradigm sells were populated (this includes all pluralia tantum and all singularia tantum nouns as sell as nouns with paradigm gaps). After this, we were left with 11320 nouns, which amounts to 158480 forms. As one can see in Table 1, there are a lot of syncretic forms in the nominal paradigms, and this holds true for every paradigm type. In our dataset, these are 89738 forms, or 56,63% of the total number of forms. This leaves 68742 (43.37%) non-syncretic forms. This indicates that syncretism is a huge difficulty for learning of Russian morphology.

Case/Number	kniga	mama	čaj	slon	yabloko	mol’
	f, inan	f, anim	m, inan	m, anim	n, inan	f, anim
	book	**mother**	**table**	**elephant**	**apple**	**moth**
Singular						
Nominative	kniga	mama	čaj	slon	yabloko	mol’
Genitive	knigi	mamy	čaja	slona	yabloka	moli
Dative	knige	mame	čaju	slonu	yabloku	moli
Accusative	knigu	mamu	čaj	slona	yabloko	mol’
Instrumental	knigoj	mamoj	čajem	slonom	yablokom	mol’ju
Locative	knige	mame	čae	slone	yabloke	moli
Genitive 2	knigi	mamy	čaju	slona	yabloka	moli
Plural						
Nominative	knigi	mamy	čai	slony	yabloki	moli
Genitive	knig	mam	čaëv	slonov	yablok	molej
Dative	knigam	mamam	čajam	slonam	yablokam	moljam
Accusative	knigi	mam	čai	slonov	yabloki	molej
Instrumental	knigami	mamami	čajami	slonami	yablokami	moljami
Locative	knigax	mamax	čajax	slonax	yablokax	moljax
Genitive 2	knig	mam	čaëv	slonov	yablok	molej

Table 1: Nominal paradigms of Russian feminine inanimate (book), feminine animate (mother), masculine inanimate (tea), masculine animate (elephant), neuter inanimate (apple) and feminine animate of a different type (moth) nouns, annotated for case and number.

2.2. Word vectors

We created our own FastText vectors by training on a cleaned version of Russian Wikipedia using the cbow algorithm and otherwise standard settings. The obtained model provides vectors with 300 dimensions and is used in the visualizations and classification experiments presented in the following sections.

3. Visualising semantic space

In this section we present the main results of exploring the data through dimensionality reduction and visualization using principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE, van der Maaten and Hinton 2008). In all of the following, the vectors of 300 dimensions were first reduced to 50 dimensions with PCA and then further reduced to two dimensions with t-SNE.

With the help of the visualizations, we explore three modifications that can help unveil the morphosyntactic functions of the vectors representing word forms: a comparison between the original and difference vectors, restricting the analysis to high frequency items, and varying the base for the computation of the difference vectors. We then see how these modifications affect the visualizations of both syncretic and non-syncretic forms.

3.1. Original vectors compared to difference vectors

The top part of Figure 1 visualizes the reduced original vectors of all the noun forms. For a better representation of syncretism and clarity, we label not individual forms, but all possible combinations of functions that can be expressed in one form. The legend of 1 is for both plots. Interestingly, although for fourteen forms 2^{14}-1 (16383) combinations are theoretically possible, only 40 are attested. We will call these combinations *case-number subsets*. The reduced semantic space has some discernible clusters, which means that some morphosyntactic functions occupy different areas in the semantic space. For example, the pink area on the top right (ablative plural words). At the same time, a lot of areas contain a mixture of vectors representing various case-number subsets.

Since it has been proposed in the literature to investigate the properties of vectors by studying shift vectors (these are vectors that represent the difference between two different forms, for example the difference between a plural and a singular; see Nikolaev et al. 2022a, Shafaei-Bajestan et al. 2022 for discussion), the bottom part of Figure 1 represents reduced difference vectors: for each noun form, a difference vector is obtained by subtracting a base vector from the vector for this form. In this figure the mean vector across all the forms of the paradigm is taken as a base vector.

The comparison of the original (top) and the difference (bottom) vectors reveals that the clusters based on the difference vectors are much clearer.

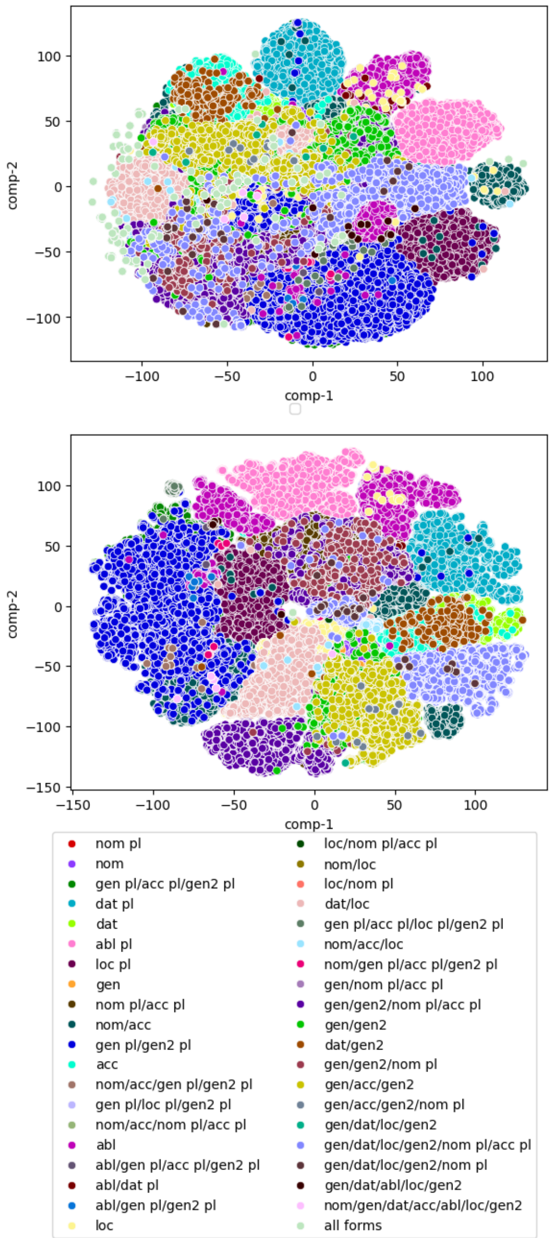


Figure 1: Visualization of nominal paradigms in Russian including syncretic ones. Original vectors are visualised on the top plot and difference vectors with the mean as a base vector on the bottom plot.

Syncretism, however, clutters the semantic space, as is to be expected. The large number of case-number subsets makes visual analysis complicated and contributes to overlapping placement of vectors belonging to different subsets. In order to be able to further reduce the effect of syncretism, we have to investigate the properties of vectors of non-syncretic items.

3.2. Removing the syncretism

In order to explore the role of syncretism, we have removed syncretic forms from the analysis. The remaining 68742 forms are visualised as the original vectors (top) as well as difference vectors (bottom) in Figure 2. We observe clearer clusters for both the original and the difference vectors as well as the separation of the same case-number representations into multiple clusters.

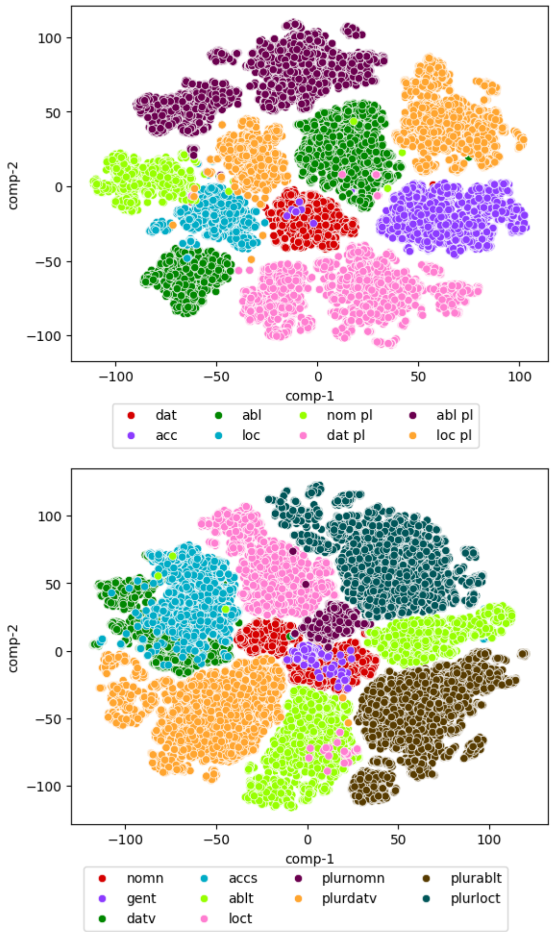


Figure 2: Visualization of nominal paradigms in Russian, only forms that are not syncretic, difference vectors. As a base, the nominative singular on top and the mean vector on the bottom.

3.3. The effect of frequency

A further idea that can be used to improve visualizations and later to approach the unsupervised learning is the option of limiting the input vectors to vectors of high frequency nouns. The reason for this is that forms of high frequency nouns occur more often in the data and thus are likely to have better representations. In addition, for high frequency lexemes we expect more forms to be encountered in the data and therefore have been learned by the algorithm and not constructed out of the n-gram representations.

The application of this approach is illustrated in Figure 3. The vectors for this visualization have

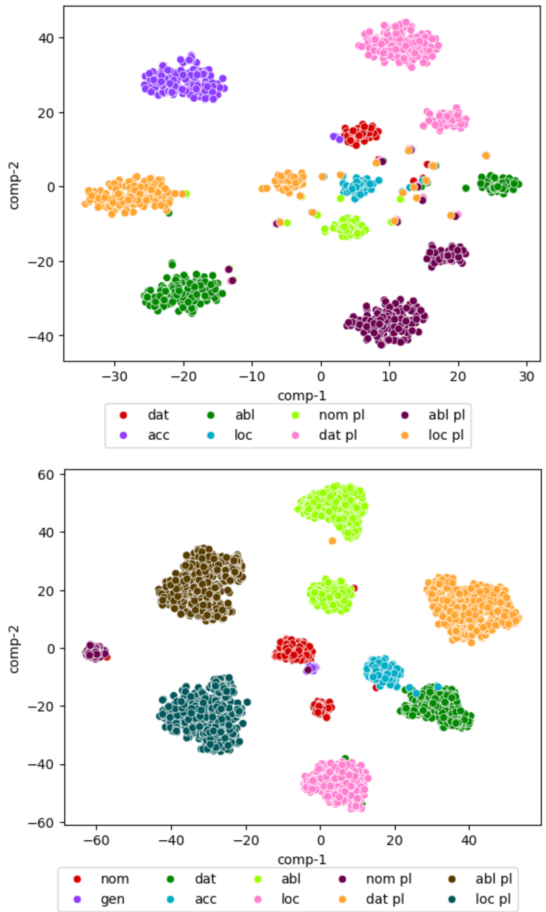


Figure 3: The syncretic forms were taken out by using dictionary information and the frequency restricted to 100 ipm or higher. Nominative singular as the base on the top and mean vector as the base on the bottom.

been filtered by frequency: only nouns with ipm of 100 and more according to Ljaševskaja and Šarov (2009). As can be seen, such restriction leads to more visually separated clusters of reduced non-syncretic vectors. Since frequency information can also be extracted during the learning process, such reduction of data can be useful in some steps of the pipeline for discovering morphological functions and their representations in a self-supervised manner.

4. Representing a lexeme

The given dictionary form is nominative singular, but is that form a good representation for the entire lexeme? For phonological reasons dictionaries also contain principal parts, which can be used to predict the phonological forms of other words in the paradigm (Albright, 2010; Nikolaev et al., 2022a). From the point of view of the embeddings, rather than the phonology of the forms, it is unclear whether principal parts are also needed to predict the semantics of other forms in the paradigm. We hypothesize that the mean vector over all forms of one lexeme represents the meaning of the lexeme better than a vector of any single form. It allows more uniform representations of morphological functions as shift vectors as well as the presence of a representation for all of the functions.

From the comparison of the top and the bottom plots in Figure 3 we can see that difference vectors are more useful when it comes to identifying clusters of vectors representing the same function, since on the bottom figure there are less morphosyntactic functions which are broken down into two distinct clusters. Let us now, after seeing the effect of frequency filtering, reintroduce syncretism, as is done in Figure 4. As above, we are considering the original vectors (top) and the difference vectors (bottom). Certain groups of forms are clearly identifiable on both visualizations: these are plural ablative or plural dative. The reason for this is that they are never syncretic (apart from the nouns that do not change their form at all, such as *mango* 'mango', absent on 4 due to the frequency restriction). As for the other number-case combinations, a couple of them form clusters without clearly identifiable borders, and most of them are spread out as in the top of figure 4. At the same time, although not all of the functional clusters can be clearly identified in the bottom of figure 4, the number of such clusters is significantly higher.

Interestingly, difference vectors representing a set of forms are often split into several clear clusters, which turn out to be related to different genders, as illustrated by 5. The most prominent example is the syncretic nominative/accusative singular form: in Figure 4 it is visible as three separate clusters. In 5 we see that these clusters correspond to the three genders, the neutral being positioned at the left periphery, the feminine between the feminine non-syncretic nominative and accusative and the masculine close to a (smaller and thus less visible) group of non-syncretic masculine nominative singular representations.

5. Supervised Classification

Although we work towards self-supervised learning, we ran a supervised classification task based on our vectors paired with morphological information. This had two goals: first, show that the data contained in the vector representations is enough in order to find all the 40 case-number sets, second to test the effect of proposed vector modifications. We built a Support Vector Machine to investigate to what extent the vectors are able to correctly use the meaning of different word forms to classify them. The words were split into 80% for training and 20% for testing. The accuracy of classification of all

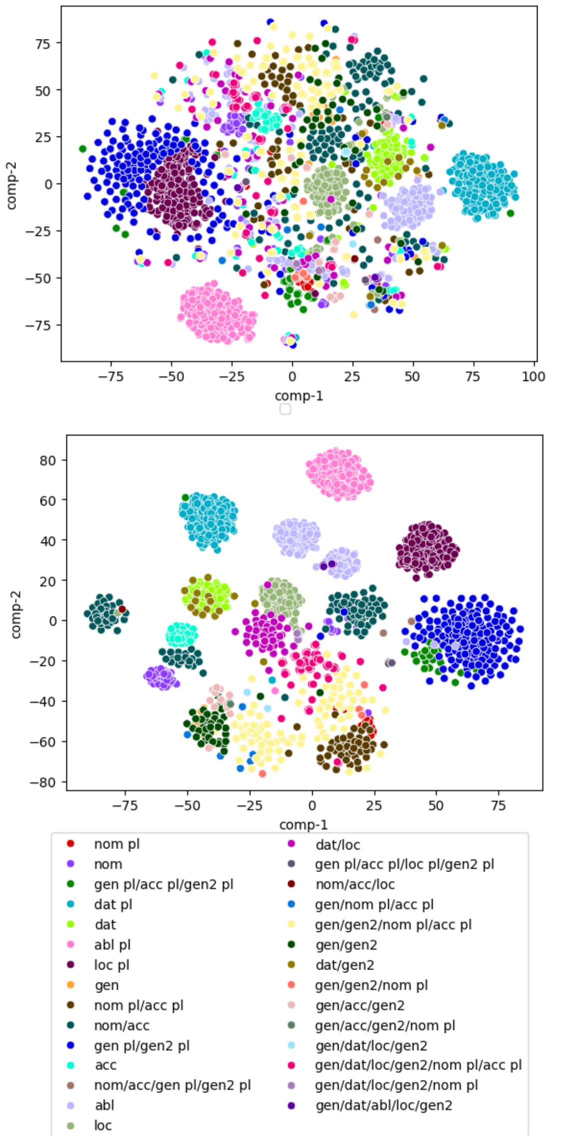


Figure 4: Original (top) and difference (bottom) vectors of relatively frequent Russian nouns including syncretic forms.

words (with syncretic forms) is 86.75%. As mentioned above, the classification needs to be done into one of the 40 categories corresponding to case-number set. This means that a word form is considered to be classified correctly only when the exact combination of morphosyntactic information is identified. So if a form is syncretic between nominative and accusative singular, it is correctly classified only if exactly this set of features is identified. This task is significantly harder than identifying one function from a set of functions a noun form can refer to.

We ran the same classification task for the subset of vectors representing non-syncretic forms. In this case the accuracy of the classification reaches 97.75%.

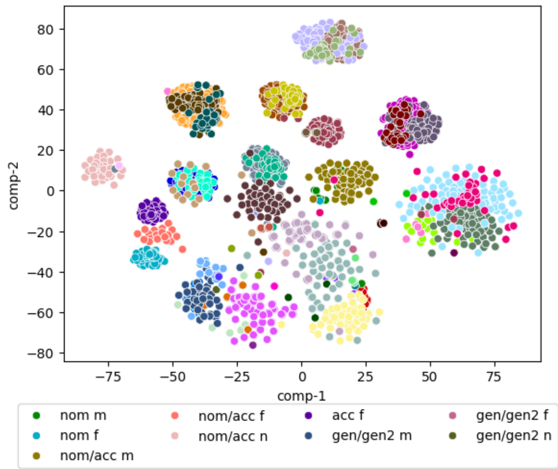


Figure 5: Difference vectors of relatively frequent Russian nouns including syncretic forms, colored by form and gender.

We have repeated both tasks for difference vectors instead of the original vectors. In both cases we see an improvement of classification: for all the forms the accuracy is 89.1% and for the non-syncretic forms it is 99.11% after this modification.

6. Semantic Classes

It is not surprising that the vector representations of nouns contain information about the semantic categories to which they belong. As we have seen, though, this information becomes less prominent if we perform a dimensionality reduction on the group of vectors that contain various case-number forms. To test the hypothesis that shift vectors correlate with the semantic class, we manually annotated 1576 nouns with 64 category labels, allowing each noun to receive multiple labels. Figure 6 represents nouns from 27 categories that are the most populous. It illustrates that if we run the PCA-tSNE reduction on the set of forms that are associated with one specific case-number function, we can observe the semantic grouping of the nouns, as illustrated by Figure 6. This result is in line with the findings of Shafaei-Bajestan et al. (2022). On the other hand, Figure 6 shows that there are no clear borders between the semantic categories and many of them get split into smaller clusters depending on their grammatical gender, as can be seen in Figure 7.

Separation between different classes becomes easier to follow if we restrict the number of classes included in the analysis, for example, to three classes that are expected to have distinct semantics, as shown in Figure 8.

Shafaei-Bajestan et al. 2022 have shown for English that the plural shift vector is not uniform across

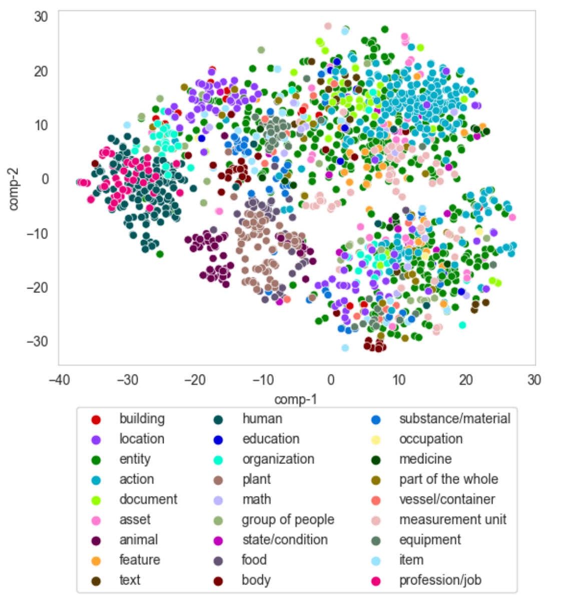


Figure 6: Original vectors of all semantic groups in dative singular

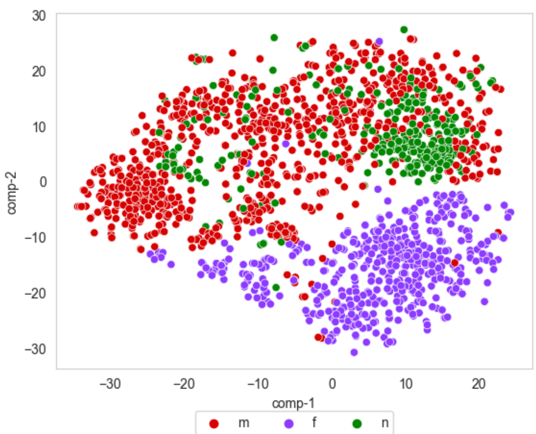


Figure 7: Original vectors of all semantic groups in dative singular colored by gender

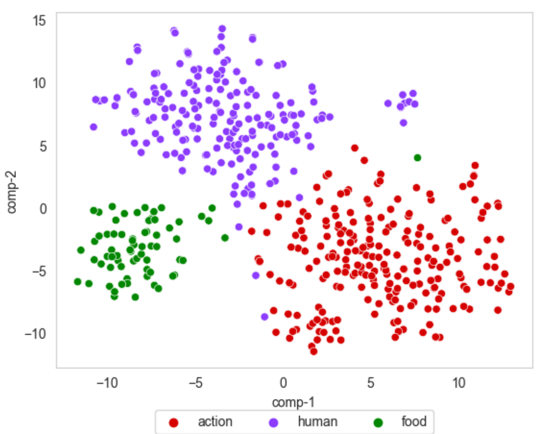


Figure 8: Original vectors of three big semantic groups in prepositional singular

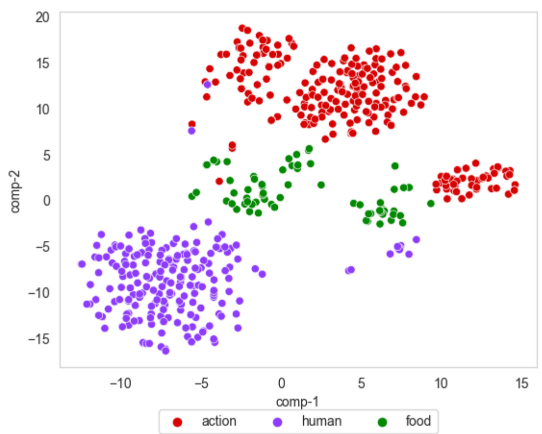


Figure 9: Shift vectors of three big semantic groups for dative singular and nominative singular as the base

semantic categories. One possible idealization would be to assume that an individual vector for a *noun* in some *form* would be a sum of the vector representation for the *noun* and the vector representation of the *form* (mutually independent). A candidate for the representation of the *noun* is the *nominative singular* representation of that noun if one assumes that nominative singular is an unmarked form and other forms are derived from it. This is the assumption in Shafaei-Bajestan et al. 2022, where the analyzed difference vectors are the difference vectors between the plural and the singular forms of a given noun. As in Russian there are many potential candidate forms when it comes to calculating a difference vector, we have explored all the possibilities of taking any given form as a base representation as well as taking the mean of all the forms of one lexeme as a base representation of the meaning of that lexeme. Our experiments have revealed that for any of the mentioned choices of the base form representation, the resulting difference vectors still carry semantic information about the class the noun belongs to. Several examples (with only three big semantic groups) are presented in 8 (original vectors) This allows us to conclude that morphosyntactic functions as learned by FastText correlate with the semantic class of the noun.

A comparison between the shift vectors with a nominative singular base, as in Figure 9, and the shift vectors with a mean base, as in Figure 10, show that the interaction between the semantic class and the morphosyntactic functions is present independently of the choice of the base form.

7. Updating representations

Based on the insights obtained from data visualisation and classifications, we have created base vector representations for nouns as well as rep-

	nom	gen	dat	acc	abl	loc	gen2	nom pl	dat pl	abl pl	loc pl	mean
all items	0.66	0.72	0.66	0.66	0.66	0.66	0.73	0.69	0.62	0.64	0.63	**0.78**
non-syncr	0.67	0.68	0.69	0.66	0.72	0.68	0.68	0.68	0.73	0.75	0.72	**0.80**
non-syncr in vocab	0.71	0.71	0.70	0.69	0.75	0.70	0.71	0.70	0.70	0.73	0.69	**0.82**

Table 2: Cosine similarity of constructed vectors with various base selection and fastText vectors

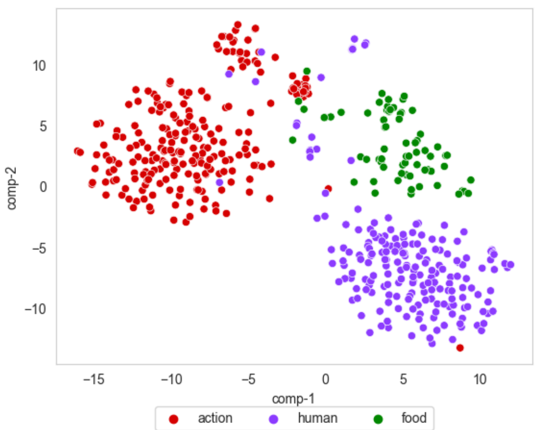


Figure 10: Shift vectors of three big semantic groups for dative singular and mean as the base

resentations for various functions and compared them with representations obtained thorough fastText. We then have compared how similar the resulting representations are to the original vectors in multiple conditions. First, we have tested every case as a potential base. For each base selection we have calculated the mean cosine similarity of all the items (first row of Table 2), the mean cosine similarity for non-syncretic items (second row) and the mean cosine similarity of non-syncretic items that are in vocabulary of the fastText model (last row). As is evident from the table, the mean as the base provided the best results, despite the fact that for each single case as a base all the items of that case would be identical to the original vectors, contributing similarity score of 1. Among the row comparisons we see that removing syncretic items and limiting the comparison to in vocabulary items increases the similarity. Based on the last observation we expect that replacing out of vocabulary representation with our constructed vectors will improve the performance of the model in downstream tasks.

8. Conclusion

We set out to investigate the effects of syncretism on learning Russian nominal paradigms from their embeddings. We are interested in doing this as a first step towards unsupervised learning of morphology. For this a pipeline has been proposed in Wiemerslage et al. (2021), but this pipeline did not take into account the effect of syncretism, which is very prevalent in Russian (our data set contained 43.4% non-syncretic forms).

We found several possible interventions that can be integrated into pipelines for semi-supervised or unsupervised learning of morphology. First, shift vectors provide a better basis for an analysis than original vectors, which is confirmed both by the visual analysis and the classification task results. The best choice of a base vector for obtaining the shift vectors, according to our observations, is an average vector of the paradigm. Since learning pipelines usually include a step of gathering forms of one paradigm, creating an average vector in an unsupervised manner should not cost additional problems and we hypothesize that in the absence of the labeled data this is the most robust choice.

We found that using high frequency items is beneficial for discovering structure in the data, both with and without syncretism. As is evident from the visual representations, in the latter case this modification is even more important and might help in the initial steps of the pipelines for unsupervised morphological learning. Although in this paper we still rely on labeled data for exploring the effect of the proposed vector modifications, we aim to leverage linguistic insights about morphological phenomena and use the resulting information to contribute to unsupervised learning of morphology.

9. Bibliographical References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

Adam Albright. 2010. Inflectional paradigms have bases too: Arguments from Yiddish. *Natural Language & Linguistic Theory*, 28(3):475–537.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Naranjo Matías Guzmán. 2020. Analogy, complexity and predictability in the russian nominal inflection system. *Morphology*, pages 1–44.

Jordan Kodner. 2022. Computational models of morphological learning. In *Oxford Research Encyclopedia of Linguistics*.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers 4*, pages 320–332. Springer.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Olga Nikolaevna Ljaševskaja and Sergej Aleksandrovič Šarov. 2009. *Častotnyj slovar sovremennogo russkogo jazyka na materialax Nacionalnogo korpusa russkogo jazyka*. Obščestvo s ograničennoj otvetstvennostju" Izdatelskij centr" Azbukovnik".

Robert Malouf. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Jessica Nieder, Fabian Tomaschek, Enum Cohrs, and Ruben van de Vijver. 2021a. Modelling Maltese noun plural classes without morphemes. *Language, Cognition and Neuroscience*, 0(0):1–22.

Jessica Nieder, Ruben van de Vijver, and Holger Mitterer. 2021b. Knowledge of Maltese singular–plural mappings: Analogy explains it best. *Morphology*, pages 1–24.

Jessica Nieder, Ruben van de Vijver, and Holger Mitterer. 2021c. Priming Maltese plurals: Representation of sound and broken plurals in the mental lexicon. *The Mental Lexicon*, 16(1):69–97.

Alexandre Nikolaev, Yu-Ying Chuang, and R. Harald Baayen. 2022a. A generating model for finnish nominal inflection using distributional semantics. *Mental Lexicon*, 17(3):368 – 394.

Alexandre Nikolaev, Yu-Ying Chuang, and RH Baayen. 2022b. A generating model for finnish nominal inflection using distributional semantics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Elnaz Shafaei-Bajestan, Masoumeh Moradipour-Tari, Peter Uhrig, and R Harald Baayen. 2022. Semantic properties of English nominal pluralization: Insights from word embeddings. *arXiv*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Chris Westbury and Geoff Hollis. 2019. Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behavior Research Methods*, 51(3):1371–1398.

Adam Wiemerslage, Arya D McCarthy, Alexander Erdmann, Garrett Nicolai, Manex Agirrezabal, Miikka Silfverberg, Mans Hulden, and Katharina Kann. 2021. Findings of the sigmorphon 2021 shared task on unsupervised morphological paradigm clustering. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 72–81.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya D. McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next.

Listen, Repeat, Decide: Investigating Pronunciation Variation in Spoken Word Recognition among Russian Speakers

Vladislav Zubov, Elena Riekhakaynen
St Petersburg University
St Petersburg, Russia
Vladzubov21@gmaul.com, e.riehakajnen@spbu.ru

Abstract

Variability is one of the important features of natural speech and a challenge for spoken word recognition models and automatic speech recognition systems. We conducted two preliminary experiments aimed at finding out whether native Russian speakers regard differently certain types of pronunciation variation when the variants are equally possible according to orthoepic norms. In the first experiment, the participants had to repeat the words with three different types of pronunciation variability. In the second experiment, we focused on the assessment of words with variable and only one standard stress. Our results support the hypothesis that listeners pay the most attention to words with variable stress, less to the variability of soft and hard consonants, and even less to the presence / absence of /j/. Assessing the correct pronunciation of words with variable stress takes significantly more time than assessing words which have only one correct pronunciation variant. These preliminary results show that pronunciation variants can provide new evidence on how a listener accesses the mental lexicon during natural speech processing and chooses among the variants stored in it.

Keywords: pronunciation variants; spoken word recognition; Russian

1. Introduction

Spoken word recognition (SWR) studies quite often address the problem of variability in the speech signal, since variability (or variation) is one of the important features of natural speech (Brouwer, 2010) and also a challenge for both spoken word recognition models and automatic speech recognition systems (Luce, McLennan, 2005). The variability of a speech signal can include individual characteristics of the speaker (timbre, dialect, accent, etc.), emotional state (tempo, intonation), speech style (formal or informal, etc.), features of the communication environment (noise and interference), see (Pufahl, Samuel, 2014) for a review. Particular attention is paid to the pronunciation variation: duration and quality, sound changes, reduction, stress. Pinnow et al. (2017) provided an example of how spoken variants can be used to assess different approaches to how a listener accesses words: either there is a set of different variants in the lexicon, or information is available in the speech signal that allows a successful comparison between the surface form and the canonical form, the latter being stored in the mental lexicon.

In particular, the paper examines reduced words and analyzes their role in the activation of unreduced canonical forms. Reduction in general is most often in the scope of the studies on pronunciation variability (see (Tucker, Ernestus, 2016)). Another type of variability is discussed by Cutler and Jesse (2021), who suggest that the stress patterns should be represented in the mental lexicon of a particular language and play a role in the process of spoken word recognition. Stress can serve as an important marker in the process of lexical access, determining which lexical items are activated in the native speaker's mental lexicon. Thus, we assume that the use of words with variable pronunciation as material for research in the field of SWR will provide new data on the lexical access and on the organization of the mental lexicon.

Despite a significant number of studies of variable pronunciation and the mental lexicon across various languages, researchers often encounter a challenge as pronunciation variants may influenced by sociolinguistic parameters. These variants can belong to different dialects, age groups, or hold varying degrees of prestige. Such characteristics impose limitations on research, as illustrated by Warren and Hay (2006).

Based on the Russian language material, descriptive studies of variation are usually carried out within the framework of orthoepy and sociolinguistics. Many papers provide rich data on modern pronunciation norms and sociocultural factors of speakers that influence the choice between pronunciation options (Kalenchuk, Savinov, 2021). However, until now, perceptual studies of pronunciation variants have not been systematically carried out. At the same time, in our opinion, the Russian language is a promising source of data on the processing of variability during SWR, since unfixed stress and active lexical processes associated with borrowing words result in numerous items with different pronunciation variants. Particularly interesting are the cases when pronunciation variants are noted by researchers as equal, i.e. there is no evidence for significant factors influencing the choice of a certain variant (context, frequency, style of speech, social status of the speaker, etc.). Thus, two or more pronunciation variants of a word exist in parallel in everyday speech and are used without any restrictions, e.g. variation of stress patterns (núzhny or nuzhný 'are needed') or variation of the consonant before the following vowel /e/ (soft or hard) ([sʲérvʲis] or [sérvʲis] 'service'), and so on. We assume that such variants can be useful for studying the ways a listener accesses the mental lexicon during natural speech processing and chooses among the variants stored in it. As far as we can conclude from the literature, such equally possible variants are not frequent in other languages. Thus, Russian data can provide quite rare evidence

The Workshop on Cognitive Aspects of the Lexicon, pages 129–132
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

on how a listener processes variation not influenced by sociolinguistic factors.
In this paper, we describe two preliminary experiments that we conducted to answer the following questions:
(1) To what extent do listeners generally notice variability in the speaker's speech, and does this depend on the variability type (different stress, substitution of sounds, or changes in the number of sounds)?
2) Is it possible to equate access to words that have several pronunciation variants with access to words that have one pronunciation variant, but are pronounced correctly or incorrectly?
The answer to the first question is explicitly stated in a few Russian-language papers (Pozharitskaya, 2004; Kasatkin, 2011; Kalenchuk, Savinov, 2021), which show that listeners pay attention to the place of stress much more often than to the segmental structure of words, but these assumptions are not supported by any experimental data. In our paper, we report Experiment 1, which offers empirical support for this proposition.
As for the second question, it is necessary to carry out preliminary studies to describe the mechanism of the recognition of words with incorrect pronunciation, and then compare these results with data obtained on the material of words with variable pronunciation. In Section 3 of the paper, we describe a pilot Experiment 2, which will be the beginning of such work.

Both experiments were conducted in the accordance with the Declaration of Helsinki and the existing Russian and international regulations concerning ethics in research.

2. Experiment 1

2.1 Method

As the goal of the experiment was to find out whether listeners pay attention to how the words with equally possible pronunciation variants are realized, we decided to ask participants to repeat the phrases they heard. There are at least two types of repetition tasks, one being the shadowing and the other – the imitation task. In the former, the participants are not given any special instructions on how accurate their repetition should be, whereas in the latter they are "explicitly instructed to imitate the productions they were exposed to" (Dufour, Nguyen, 2013). Dufour and Nguyen (2013) have shown that the general mechanism revealed by these two experimental paradigms is probably the same and provides evidence on how the words are stored in the long-term memory. Thus, we instructed our participants that they should just repeat what they heard. We supposed to obtain the information on how accurately participants process different types of pronunciation variation.

2.2 Stimuli

We chose the material for the experiment from the Big Orthoepic Dictionary of the Russian Language (https://gramota.ru/biblioteka/slovari/bolshoj-orfoepicheskij-slovar-russkogo-yazyka). According to it, all the words we used in the stimuli can have two pronunciation variants and these variants do not depend on the age and other parameters of the speakers and are considered equally appropriate to be used by the native speakers of Russian. We compared three types of variation: 1) Stress: variation of stress patterns (e.g. núzhny or nuzhný 'are needed'); 2) CV: variation of the quality of a consonant before the following vowel /e/ (soft or hard) ([sʲérvʲis] or [sérvʲis] 'service'); 3) VJV: presence or absence of the consonant [j] between two vowels (proekt [proekt] or [project] 'project'). For each group, we chose 12 words. These were mainly nouns (26 out of 36), but also five adjectives, four verbs and one adverb. Nouns are the most frequent words in the Russian language, and it seems that the phonetic variation of the three types we chose for our study occurs in these words more often than in other parts of speech. We included in the experiment 12 fillers (the words without pronunciation variants) which were also mainly nouns.
We created two-word constructions with all the words, which were read by a male speaker and audio-recorded. For all the stimuli, we recorded both pronunciation variants; fillers were recorded only once, as they had only one possible pronunciation. Then, we arranged all words into two stimuli lists. Each list included 12 fillers and one of the two possible pronunciation variants for each of 36 stimuli. The duration of both stimuli lists was about 3.5 minutes.

2.3 Procedure

During the experiment, participants listened to one of the two audio recordings via headphones and were asked to repeat after the speaker exactly what they heard. They were given 3 seconds to respond to each stimulus. The experimenter documented whether the variant pronounced by the participant matched the one in the recording.

2.4 Participants

96 native speakers of Russian took part in the experiment (62 female; M_{age} = 19 y.o.). None of them reported any hearing problems. All participants provided an oral consent to take part in the experiment.

2.5 Results

The number of correct repetitions (CORR) after the speaker for each individual stimulus was analyzed (regardless of the pronunciation variants, since the number of their presentations was equal). The mean CORR (Max = 96) and standard deviation (SD) for each type are provided in Table 1.

Variation type	CORR Mean	SD
VJV	51.75	5.29
CV	66.08	8.21
Stress	85.75	6.65

Table 1: Average correct repetitions and standard deviations for each condition

The smallest number of correct repetitions was in the group 3 VJV (with the presence or absence of the intervocalic /j/) – 53.9%, and the largest – in the group of words with variable stress (89.3%). To test whether the differences were significant, a linear regression model was fit. The outcome variable was CORR, and the predictors were the type of variation, which had three levels: VJV, CV, Stress (see Table 2).

	Estimate	SE	t	p
(Intercept)	54.85	1.91	28.73	< .001
CV	15.95	2.70	5.91	< .001
Stress	38.10	2.70	14.11	< .001

Table 2: Summary of significant effects in the number of correct repetitions

Neither the frequency of word forms of the selected words, nor the part of speech had a significant effect on the number of correct repetitions after the speaker, and thus these parameters were not included in the model. It can be concluded that the number of correct repetitions strongly depends on the type of variability. Listeners pay the most attention to words with variable stress, which was noted in previous papers (Pozharitskaya, 2004; Kasatkin, 2011; Kalenchuk, Savinov, 2021). The change in the quality of the consonant sound before the vowel is less prominent for the native speakers of Russian, whereas the presence or absence of an intervocalic /j/, apparently, is not noticed in speech, since the number of correct repetitions behind the speaker is close to random. In the next experiment, we decided to test how lexical access to words with variable stress occurs.

3. Experiment 2

3.1 Method

Reaction time is a measure which is commonly used to study lexical access. Most often the reaction time is measured while participants perform a lexical decision task (LDT). As in our study we focus on pronunciation variation, we measured reaction time while participants were deciding whether the given word is correct or not. Thus, we used a modified version of the LDT.

3.2 Stimuli

We recorded 30 isolated words for the experiment. Their pronunciation was checked in the same orthoepic dictionary as in the first experiment. There were three groups of words: 1) those with variable stress (for each of them we recorded two stimuli with both variants); 2) with the only one standard stress and pronounced correctly by the speaker; 2) with the only one standard stress but pronounced incorrectly by the speaker.

3.3 Procedure

The experiment was conducted in PsychoPy. Each participant was presented with 30 isolated words in random order through headphones; one of the two possible stimuli for every word with variable stress was chosen randomly by the program. After listening to each stimulus, participants answered whether the word sounded correct or not by pushing one of two buttons on the computer. Reaction time (from the beginning of listening to making a decision) and the correctness of answers to questions were measured.

3.4 Participants

25 people took part in the experiment (20 female; Mage = 18 y.o.). None of them reported any hearing problems. All participants provided an oral consent to take part in the experiment.

3.5 Results

We analyzed the average reaction time (RT, ms) in each of the stimulus groups (747 reactions in total), as well as the answers of the participants (in which cases the stimulus was considered correct, the percentage of the total number of responses).

Pronunciation	RT (ms)	SD	Answers “correct”
Variable	2248.69	859.31	64.7%
Incorrect	2090.45	786.12	6.4%
Correct	1592.17	509.68	100%

Table 3: Average mean RT, standard deviations and the percentage of the answers “correct” for each condition

The Table 3 shows that words with variable stress are rated as correctly pronounced in 64.7% of all cases, while the words with the only one correct stress (group 2) are rated as correct by all participants and the incorrectly pronounced stimuli are most often considered incorrect.

Words with variable stress required the greatest amount of time for participants to react, but we should note that the standard deviation in this group is the largest.

To assess the statistical significance of the results obtained, we used linear regression with RT as the dependent variable. The group of stimuli, in which there were three levels (variable, irregular and correct) and the frequency of word forms according to the Russian National Corpus (https://ruscorpora.ru/en/) were used as predictors (Table 4). Length of stimuli in number of sounds and part of speech did not show a statistically significant effect and were not included in the model.

	Estimate	SE	t	p
(Intercept)	2219.75	47.97	46.27	< .001
Incorrect	-99.29	68.14	-1.46	0.146
Correct	-562.00	70.93	-7.92	< .001
Freq (log10)	-67.37	20.40	-3.30	0.001

Table 4. Summary of effects in RT

The word form frequency plays a role in the evaluation of stimuli, even though a modified LDT technique is used, so this factor needs to be considered in future studies.

The RT for the group of words with regular stress turned out to be statistically significantly lower compared to the other groups, but no statistical difference was achieved between the groups of stimuli with irregular and variable stress. It is assumed that the lack of a statistically significant difference may be due to the heterogeneity of the stimuli in these groups, since it is not easy to select words of the same length, part of speech and frequency for the Russian language, because we lack a database of words with pronunciation variants.

4. Discussion and Conclusion

In this study, we expand on the concept of variability in the Russian language from a perceptual perspective and try to assess the role of pronunciation variability in the process of SWR. Based on the results of two experiments, we can conclude that, firstly, listeners notice variability in speech in different ways, and when repeating after the speaker, in some cases they activate the same units that they heard, and in the others – those that are stored in their mental lexicon and not necessarily matching the heard variant.

We provided experimental evidence for the assumption of Russian orthoepy experts that a naive native speaker of Russian, when assessing pronunciation, pays more attention to word stress, less to the variability of soft and hard consonants, and even less to the presence / absence of /j/ (Pozharitskaya, 2004; Kasatkin, 2011; Kalenchuk, Savinov, 2021). Secondly, assessing the correct pronunciation of words with variability takes significantly more time than assessing words which have only one correct pronunciation variant. However, it is not clear how the process of accessing words with variable pronunciation occurs: whether it is similar to how words with incorrect stress are recognized or differs from it. We hypothesize that further exploration of the variability phenomenon in Russian from a perceptual perspective will yield insights into these questions.

The limitations of the current study include the following:

a) the level of conducted experiments is rather shallow, since the results do not allow us to draw conclusions about the access to the listener's mental lexicon. However, the results obtained show the promise of further research into the described language material;

b) the sets of stimuli for both experiments were unbalances because of the absence of a database containing pronunciation variants in Russian;

c) it is necessary to compare the results with similar data from other languages, in which we can find equally possible pronunciation variants;

d) the documentation of accurate repetitions after the speaker in the first experiment relying on the experimenter's hearing might have influenced the results (particularly in the VJV group).

In our further research on Russian, we intend to conduct a more careful selection of stimuli. This selection will allow to include various factors into the model (frequency, morphological features, morphemic composition, etc.). We also plan to develop designs for more complex experiments aimed at gathering data on the process of SWR.

5. Acknowledgements

We thank Elizaveta Zhukatinskaya and the students who took part in the linguistic program in the Educational Centre "Sirius" (Sochi, Russia) in October 2021 for the help with collecting and analyzing the data.

6. Bibliographical References

Brouwer, S. (2010) *Processing strongly reduced forms in casual speech*. PhD Thesis, Radboud University Nijmegen, Nijmegen.

Cutler, A., and Jesse, A. (2021). Word Stress in Speech Perception. In J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni (Eds.), *The Handbook of Speech Perception*, Wiley, pp. 239–265). https://doi.org/10.1002/9781119184096.ch9

Dufour, S., and Nguyen, N. (2013) How much imitation is there in a shadowing task? *Frontiers in Psychology:* 4. https://doi.org/10.3389/fpsyg.2013.00346

Kalenchuk, M. L., Savinov D. M. (Eds.). (2021). Pronunciation standards in usage and codification (in Russian). Moscow: Russian Language Institute.

Kasatkin, L. L. (2011). Orthoepeme as the Basic Unit of Orthoepy. *Voprosy Jazykoznanija*, 2: 31–38.

Luce, P. A., and McLennan, C. T. (2005). Spoken Word Recognition: The Challenge of Variation. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception*, Blackwell Publishing Ltd. pp. 590–609. https://doi.org/10.1002/9780470757024.ch24

Pinnow, E., Connine, C. M., and Ranbom, L. J. (2017). Processing pronunciation variants: The role of probabilistic knowledge about lexical form and segmental co-occurrence. *Journal of Cognitive Psychology*, 29(4): 393–403. https://doi.org/10.1080/20445911.2017.1279619

Pozharitskaya, S. K. (2004). Orthoepy: Ideas and Practices. In G. E. Kedrova, Potapova V. V. (Eds.) *Language and Speech: Problems and Solutions*. Moscow: MAKS.

Pufahl, A., and Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology*, 70: 1–30. https://doi.org/10.1016/j.cogpsych.2014.01.001

Tucker, B., and Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon*, 11(3): 375-400. https://doi.org/10.1075/ml.11.3.03tuc

Warren, P., and Hay, J. (2006). Using sound change to explore the mental lexicon. In. C. M. Flinn-Fletcher & G. M. Haberman (Eds.), *Cognition and Language: Perspectives from New Zealand*, Australian Academic Press. pp. 105-125.

The Mental Lexicon of Communicative Fragments and Contours: The Remix N-gram Method

Emese K. Molnár, Andrea Dömötör
National Laboratory for Digital Heritage (DH-LAB)
Budapest, Hungary
emesekmolnar@gmail.com, domotor.andrea2@btk.elte.hu

Abstract

The classical mental lexicon models represent the lexicon as a list of words. Usage-based models describe the mental lexicon more dynamically, but they do not capture the real-time operation of speech production. In the linguistic model of Boris Gasparov, the notions of communicative fragment and contour can provide a comprehensive description of the diversity of linguistic experience. Fragments and contours form larger linguistic structures than words and they are recognized as a whole unit by speakers through their communicative profile. Fragments are prefabricated units that can be added to or merged with each other during speech production. The contours serve as templates for the utterances by combining linguistic elements on specific and abstract level. Based on this theoretical framework, our tool applies remix n-grams (combination of word forms, lemmas and POS tags) to identify similar linguistic structures in different texts that form the basic units of the mental lexicon.

Keywords: remix n-gram, communicative fragment, communicative contour, text reuse

1. Introduction

Models of the mental lexicon have long been central to linguistics and several very different approaches have been developed over time. Our paper seeks to answer the following questions: (i) What kind of mental lexicon model can capture the everyday linguistic experience of the language user and describe the language production at the discourse level? (ii) How can we operationalise the theoretical model, or in other words which (NLP) method can fit the theory?

After the structuralist concept of Saussure, the Chomskyan generative grammar (Chomsky, 1965) has dominated the field of linguistics. Based on this theory the basic unit of language is the sentence which can be built using elements and rules. The building blocks are the words that are contained in the lexicon as a list of elements. In contrast to this formal concept, the functional and usage-based approach (Langacker, 1987; Croft, 2001; Goldberg, 2006; Bybee, 2010) focuses on not just the language itself, but on the speaker as well, since these two cannot be separated. Just as the knowledge of language cannot be separated from general knowledge of the world, neither can the lexicon be separated from the rest of the language. They describe linguistic units as form and meaning pairs, holistic and emergent.

Although these lexicon models are no longer simple collections of words, as they change from static to continuum concept, they are still not dynamic enough to describe the online, i.e. real-time process of language production. Even if they focus on the speaker, these approaches still concentrate on the abstract form of the language, and fail to capture the everyday linguistic experiences of language users. Accordingly, this paper presents a theoretical model in which lexical items and grammar are emergently linked and proposes an NLP-based tool inspired by this model. We tested our method on a Hungarian poetry corpus as a first step in the development of the tool. By the end of the project, the aim is to develop a tool that can detect the typical communicative fragments and contours in different texts.

2. Communicative Fragments and Communicative Contours

Boris Gasparov built his intertextual concept of language on Bakhtinian theory. Gasparov criticizes construction grammars and Cognitive Grammar because, although they are not rule-based descriptions, their concepts are still too abstract and rigid. His central concept is the communicative fragment (CF), which is „a concrete segment of speech of any shape, meaning, and stylistic provenance that speakers are able to recognize spontaneously and to use as a conventional expression that fits certain communicative purposes" (Gasparov, 2010: 38). CFs can be more varied than constructions, since their boundaries are defined by the linguistic experience of the language user. Thus, fragments - unlike lexical units in a traditional lexicon model - are not listable but are constantly changing; they do not necessarily have either a compact or a fixed syntactic structure; they have the ability to evoke, allude to, and merge with each other (Gasparov, 2010: 50-55). Besides, CFs are prefabricated and

The Workshop on Cognitive Aspects of the Lexicon, pages 133–139
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

ready-made pieces as well that are embedded in context. Every CF has a texture that is an imprint of the specific situations in which it is used. This texture determines the expectations of discourse, i.e. the communicative profile that marks the genre, style, potential topic and conversing parties associated with the fragment (Gasparov, 2010: 55-58).

(1) Open the door to the west veranda!
[open the door]
[the door to]
[the door to the veranda]
[the west veranda]

Example (1) shows that CFs can vary in length and overlap. As Gasparov (2010) describes: „Due to their doubleedged connection to fluid mental processes on the one hand and to linguistic hardware on the other, CFs constitute the crucial link between the cognitive and the operational aspects of language – between creative efforts of the mind and the concrete material that allows those efforts to emerge as tangible facts of speech" (Gasparov, 2010: 64).

According to Gasparov (2010) CFs constitute a primary vocabulary that is at least as important for speakers' knowledge of their language as the vocabulary of lexical units. Speakers compose and interpret speech primarily based on CFs rather than words. Although a CF can be divided into smaller meaningful components like words and morphemes, they remain a single unit for the speaker. Although they can recognize the components and complex structure of the unit this analytic process is not reflected during the speech production like the subprocesses are not reflected either in the habitual operations of their everyday life (Gasparov, 2010: 39).

In the Gasparovian model, CFs are considered the basic unit of language, but for speech production a speech prototype (SP) is needed that leads to the realization of the speech artifact (SA) (Gasparov, 2010: 117-121). The starting point for the SA is the fragments, which are linked together through communicative profiles. The fragments are connected to each other on the basis of both structural and semantic similarities, and the resulting networks are the speech prototypes. In addition to the SP, the realisation of the SA requires to define the specific contextual framework. While the fragment carries its specific context in its texture, the speech artifact always needs motivation based on the actual context. The essential aspect of the speech process is the way in which the prefabricated fragments are organised into prototypes, as possible variants of which the speech artifacts are created according to the specific context.

1st segment	2nd segment	3rd segment	4th segment
Actually I'm	[surprised] [amazed] [glad] [so glad] [so happy]	you/he/she/they/John	[VP].

Figure 1: Communicatice Contour

Besides, the communicative fragments, communicative contours (CC) are available for the speakers to create utterances. Similarly to a CF, CC is recognized comprehensively by the speaker and it has an imprint of the situation where it occurred. The difference between the CF and CC is the character of their shape. Contours can be seen as a template rather than a blueprint. CC is a semi-concrete design with some prefabriceted pieces and gaps between the structural elements. The gaps are flexible during the speech process, they can be contracted, expanded or reshaped in order to complete the utterance. While CFs are incomplete and fragmentary, its borders are often vague, so it can be easily modified or fused with other CFs in speech. In contrast, a CC has to be structurally complete and it has a sharply outlined frame. Its flexibility comes from within, filling of the gaps can be varied, while the structural elements retain the specific character of the CC (Gasparov, 2010: 151-159).

Accordingly, the CC is built up of three constituents: a lexical-structural template (Gasparov, 2010: 158), a prosodic template (Gasparov, 2010: 162) and lacunae (Gasparov, 2010: 166). The lexical-structural template containing morphemes, words or word combinations is considered to be the most sharpened constituent with the most concrete elements. This template includes the elements that function as signposts, which help to select and place other possible elements in key positions in the structure. In comparison, a prosodic template is less a concrete form that is not identical with the actual pronunciation of the utterance, but rather a comprehensive sound shape in the inner perception of speakers. It determines, among others, the intonational contours of pronunciation and pauses. Together, the pitch curve, rhythmical texture, accent and timbre of voice as prosodic signposts guiding the vocalisation of the CC to complete syntactic patterns as well as to select specific elements of the lexicon. The least specific constituent of the contour is the lacunae between the lexical and prosodic markers. However, these gaps cannot be considered as blank space, as they consistently fit into the overall structure of the contour. The main lexical items, the rhythmic and intonational contour and the general communicative profile delimit the set of items that can be inserted into the lacunae.

Speakers keep in memory a large number of CCs of different shapes, lengths, and styles. Similar to the vocabulary of CFs, the vocabulary of CCs does not form a coherent system. It can rather be described as a shapeless agglomeration of templates in speakers' memory. Different CCs are linked to different communicative situations and speech experiences, each activated in a speaker's mind opportunely. They are package of knowledge whose relation to each other, and to a presumable overall system, is simply irrelevant to speakers (Gasparov, 2010: 157).

3. Linguistic Remix

Gasparov built his theory on the basis of intertextualty. "The prevalent mode of speakers' linguistic activity can be called “intertextual,” in the sense that speakers always build something new by infusing it with their recollection of textual fragments drawn from previous instances of speech" (Gasparov, 2010: 3). His thoughts are very close to those of another author, who describe not just the language, but the whole culture in the spirit of reuse. Lessig (2008) is credited with the concept of remix culture which questioned and renegotiated not only the term of authorship, but also the understanding of creativity and culture. The concept of remix can be adapted successfully to the examination of cultural practices because – as Lessig points out – while the phenomenon of remixing may seem novel, its core mechanism has long been a part of human culture: remixing with (digital) media is identical to the fundamental process of language use. Evoking and incorporating the words of others into written works or conversations is so natural that we do not even notice the borrowing.

Popular text generation tools which are based on the Large Language Models (LLMs) owe their success to the fact that they exploit this fundamental linguistic mechanism. These tools learn from large amount of textual data sets to determine the probabilistic values of linguistic patterns for generating textual content. In other words, they remix by recognising and regenerating existing linguistic patterns. However, the output of LLMs is not completely transparent from this point of view, so it is also worth developing tools that highlight and and reveal the basic patterns of language production.

Remix can be seen as a concept that can capture both theoretical and methodological approaches based on the everyday linguistic experiences of language users. In contrast to intertextuality, it is not only suitable for describing prototypical, lexical repetitions, but can also be applied to the investigation of linguistic similarities on the structural level. The remix can serve as a framework and can link the Gasparovian language model and an NLP solution that fits the theory.

4. Related Work

There are many examples of computational methods for intertextuality and text reuse detection. These studies usually focus on the relationships between texts in different text types. For example, one line of research focuses on texts that are reused in academic work, with a particular focus on plagiarism (Citron and Ginsparg, 2015; Anson and Moskovitz, 2020, Gienapp et al., 2023). Large corpora of newspapers are available for the study of text repetition as well (Smith et al., 2013; Vesanto et al., 2017; Rosson et al., 2023). Intertextuality in literary texts is a long-established and widely researched phenomenon, which has been further enhanced by the increasing availability of large corpora and the emergence of computational methods. (Kahane and Mueller, 2001; Lee, 2007; Coffee et al., 2013; Büchler et al., 2014; Gladstone and Cooney, 2020).

The closest in spirit to our own project were those tools that make possible to detect text reuse within literary corpora. Both the Chicago Homer (Kahane and Mueller, 2001), the Tesserae (Coffee et al., 2013) and the Commonplace Cultures (Gladstone and Cooney, 2020) have a query interface with search and comparison function. Among these the Chicago Homer is a bilingual database of Early Greek epic. The corpus is tokenized, lemmatized and annotated with morphological and narratological tagging. The tool makes possible to find repetitions (sequence of two or more words) in the corpus and to filter them by various criteria. The Tesserae Project provides an online tool that allows users to compare two texts in ancient Greek, Latin, or English. The basic Tesserae search finds sentences or poetic verse lines in two different texts that share two or more similar lemmata based on an n-gram method. Experimental search options were added for sound similarity, for semantic relatedness of Greek to Latin and for context similarity using a topic modeling approach. The Commonplace Cultures project aims to detect text reuses in the Eighteenth Century Collection Online (ECCO). For the comparison of the texts PhiloLine, a sequence alignment tool was developed. The model is based on shingles of n-grams to find shared passage according to the number of common contiguous n-grams between two textual sequences.

5. Method

While text reuse research usually focuses on a single text type, our goal is to develop a method to detect typical patterns in several text type. Most of the tools are designed for English texts, so our further goal is to have a suitable tool for examining another type of language, a morphologically rich language.

In most cases, analysis of text reuse apply n-gram-based methods to find text similarities. In the case of word n-grams, the text is divided into sequences of n adjacent words in particular order. The result is similar to the example about CFs that was shown in example (1). This suggest that n-gram based methods can be suitable for detecting potential CFs within a text. By comparing the n-grams of different texts, we can identify the fragments that are usually characteristic of a discourse based on their repetition. Based on this, our method started by comparing trigrams of texts.

5.1. Remix n-gram

As we have seen in the case of CCs in figure (1), the production of utterances requires not only the combination of concrete words or fragments, but a template that are consist of linguistic units on the different levels of abstraction. Our method, called remix n-gram, is based on the concept of mosaic n-gram of (Indig and Bajzát, 2023). Mosaic n-grams are combinations of words, lemmas and POS tags representing different levels of language both specific and abstract. The comparison of such n-grams can capture structural similarities of texts besides the textual ones.

The first challenge was the sentence segmentation as we used a poetry corpus (see in detail in section 5.2). In the case of poems that contain sentence punctuation marks we segmented the text according to these. Many poems, however, does not contain sentence punctuations, therefore in these cases we considered each stanza a sentence.

As a next step we removed determiners (*a, az* "the", *egy* "a, an"), the conjunction word *és* ("and") and all the words tagged as "other" (X) by the morphological analyzer. We extracted the trigrams of the remaining words of each sentence. The trigrams contained all the information of the word: the word form, the lemma and the POS tag. For POS tag we used the UD tagset enhanced with some language-specific characteristics. This was necessary because Hungarian is a morphologically rich language therefore a UD POS tag and lemma cover many different word forms that are not really grasped as similar words by language users. Nominal tags (NOUN, ADJ, PROPN, PRON, NUM) were enhanced with case (2a) and a "Poss" feature if they were possessive (2b). In the case of adjectives we also marked the different degrees (2c). Verbs were divided into finite and non-finite (infinitive) groups (2d and 2e) and the tags of finite verbs were enhanced with the mood feature (2e).

(2) a. *erdőbe* 'to the woods'
`NOUN` → `NOUN.Ill`

b. *lelke* 'his/her soul'
`NOUN` → `NOUN.Poss`

c. *kisebb* 'smaller'
`ADJ` → `ADJ.Cmp`

d. *látni* 'to see'
`VERB` → `VERB.Inf`

e. *mennék* 'I would go'
`VERB` → `VERB.Cnd`

We compared the trigrams of each poem, taking into account word forms, lemmas and POS tags, and ranked the degree of similarity between them. The matching word forms got the highest score, followed by matching lemmas and POS tags. This means that a matching word form is worth 3 points for each token, a matching lemma is worth 2 points and a matching POS tag is worth 1 point. So, if three word forms match, the trigram is worth 3x3 points, that is 9 points, if two word forms and one lemma match, the trigram is worth 2x3+2, that is 8 points in total and so on. The minimal requirement of similarity was having all three POS tags and at least one lemma matched.

5.2. Corpus

To test our method, we chose a corpus that was available with the annotation layers to extract the remix n-grams. For this reason, we used The ELTE Poetry Corpus (Horváth et al., 2022a) of 3,441,864 tokens contains the complete poems of 50 Hungarian canonical poets. Besides, tokenization, lemmatization, as well as the part-of-speech and morphological analysis, the automatic annotation of the structural elements (title, stanzas, lines) and the sound devices (rhyme scheme, rhyme pairs, rhythm, alliteration, phonological structure of words) of the poems were completed in XML format.

6. Results

As a pilot study, we chose a poet from the poetry corpus and compared the remix trigrams of the poems among each other. This allowed us to identify potential candidates of CFs and fragments of CCs specifically, instead of the prototypical, literal cases of intertextuality. We compared 514 poems, in which we found more than 200,000 matching trigrams.

Examples (3)-(7) show different degrees of matching in the examined subcorpus. (3) is an exact match, (4) has two matching words and one matching POS tag, (5) and (6) have one matching word and two matching POS tags, and (7) has one matching lemma and two matching POS tags.

(3) *este van már – este van már*
evening is already – evening is already

'It's evening already'

(4) *soha nem látott – soha nem hallott*
never not seen – never not heard

'never seen' – 'never heard'

(5) *mint rossz madár – mint jámbor állatok*
like bad bird – like pious animals

'like a bad bird' – 'like pious animals'

(6) *mint rossz madár – mint puszta rom*
like bad bird – like mere ruin

'like a bad bird' – 'like a mere ruin'

(7) *pogányok lelke volt – emberek emléke van*
pagans soul.Poss was – people memory.Poss is

'It was pagans' souls' – 'there is a memory of people'

As the examples show the remix n-gram method is suitable for capturing structural similarities, however many of the matches on more abstract levels were not matching in terms of semantic similarity. It can be seen from the (6) example that one word and two POS tag matching does not always fulfill our expectations. While in example (7) the bird and the animal are still semantically close to each other, in the case of example (6) the bird and the ruin are not elements of similar semantic categories.

7. Discussion and future work

The aim of our method was to constract a model of the mental lexicon that correlates with the everyday linguistic experience of language users. Instead of a formal and list-like description of the lexicon, a functional and usage-based approach was adapted for this purpose. The communicative fragment and contour of the Gasparovian concept were suitable to serve as the basis for a dynamic model. The proposed remix n-gram method is effective in identifying the potentional text passages of more specific word-level CFs and sequences similar to CCs with more abstract structures. Since we extract typical language structures by comparing texts, we get different results from the comparisons of different texts, just as different language users have different language experiences. Thus, an NLP tool was provided that is theoretically grounded and the concepts of the theoretical framework became operationalized in practice as well.

We are able to find potential CFs and fragments of CCs using the remix n-gram method, but the large number of hits is still difficult to manage, meaning that the method needs to be refined in order to increase precision. We can reduce the number of hits by extending the stop word list or by further specifying the POS tags. Currently, only mood is specified as a criterion for verbs, but by adding number, person and tense could give more accurate hits. Furthermore, the (5) and (6) examples show that filtering words that are semantically closer to each other could also lead to better results. To find semantic similarities, it would be worth using word embedding to rank matches such as *bird* and *animal* over *bird* and *ruin*.

The next stage of the project will be to test the method on different corpora. Firstly, the ELTE Poetry Corpus has a folk song subcorpus (Horváth et al., 2022b), which is closer to the oral culture, and besides this the ELTE Novel Corpus (Bajzát et al., 2021) and the ELTE Drama Corpus (Szemes et al., 2022) are also available. The ongoing Lyrical Poetry Corpus project (Horváth et al., 2021) of ELTE DiAGram Research Centre for Functional Linguistics will contains not only canonical poems, but also song lyrics and slam texts in Hungarian. The Hungarian gold standard corpus project (K. Molnár and Dömötör, 2023) of DH-Lab will provide the opportunity to test less artistic text types, closer to everyday discourse, such as texts from blogs, educational and cultural web sites, in addition to novels.

In addition to testing on different corpora, we also aim to compare with other methods. For example, PhiloLine, which is used in the Commonplace Culture project, is an open source tool, so it can be used on different corpora. As it is also an n-gram based method, it offers the possibility to test the effectiveness of remix n-grams for languages with rich morphology. This may also pave the way for further developments aiming at applying the remix n-gram method to other types of languages.

These methods of identifying language structures are also useful because they bring us closer to understanding how language works in general. Unlike the tools based on LLMs, the results are more transparent and easier to interpret. In the long term, they can therefore contribute to the development of LLMs.

Acknowledgement

The research was supported by the ÚNKP-23-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the

National Research, Development and Innovation Fund.

References

Ian G. Anson and Cary Moskovitz. 2020. Text recycling in stem: a text-analytic study of recently published research articles. In *Accountability in Research 28*, pages 349–371.

Tímea Borbála Bajzát, Botond Szemes, and Eszter Szlávich. 2021. Az elte dh regénykorpusz és lehetőségei. In *Online térben – az online térért. Networkshop 30: országos online konferencia*, pages 63–72, Budapest. HUNGARNET Egyesület.

Joan Bybee. 2010. *Language, usage and cognition.* Cambridge University Press, Cambridge.

Marco Büchler, Philip R. Burns, Marco Müller, Emily Franzini, and Greta Franzini. 2014. Towards a historical text re-use detection. In *Text Mining: From Ontology Learning to Automated Text Processing Applications*, pages 221–238. Springer.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Mass.

Daniel T. Citron and Paul Ginsparg. 2015. Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences*, 112(1):25–30.

Neil Coffee, Jean-Pierre Koenig, Poornima Shakti, Roelant Ossewaarde, Chris Forstall, and Sarah Jacobson. 2013. The tesserae project: Intertextual analysis of latin poetry. *Literary and Linguistic Computing*, 28:221–228.

William Croft. 2001. *Radical Construction Grammar. Syntactic theory in typological perspective*. Oxford University Press, Oxford.

Boris Gasparov. 2010. *Speech, memory, and meaning: Intertextuality in everyday language*. De Gruyter, Berlin.

Lukas Gienapp, Wolfgang Kircheis, Bjarna Sievers, Benno Stein, and Martin Potthast. 2023. A large dataset of scientific text reuse in open-access publications. *Scientific Data*, 10(58).

Clovis. Gladstone and Charles Cooney. 2020. Opening new paths for scholarship: Algorithms to track text reuse in ecco. In *Digitizing Enlightenment: Digital Humanities and the transformation of Eighteenth-Century Studies*, pages 353–374. Voltaire Foundation in association with Liverpool University Press.

Adele Goldberg. 2006. *Constructions at work. The nature of generalization in language*. Oxford University Press, Oxford.

Péter Horváth, Péter Kundráth, Balázs Indig, Zsófia Fellegi, Eszter Szlávich, Tímea Borbála Bajzát, Zsófia Sárközi-Lindner, Bence Vida, Aslihan Karabulut, Mária Timári, and Gábor Palkó. 2022a. Elte poetry corpus: A machine annotated database of canonical hungarian poetry. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3471–3478, Paris. European Language Resources Association (ELRA).

Péter Horváth, Péter Kundráth, and Gábor Palkó. 2022b. Elte népdalkorpusz – magyar népdalok gépileg annotált adatbázisa. In *Valós térben – Az online térért: Networkshop 31: országos konferencia*, pages 276–283, Budapest. HUNGARNET Egyesület.

Péter Horváth, Gábor Simon, and Tátrai Szilárd. 2021. A lírai személyjelölés konstrukcióinak annotálási elveiről. In *Líra, poétika, diskurzus*, pages 133–166, Budapest. ELTE Eötvös Collegium.

Balázs Indig and Tímea Borbála Bajzát. 2023. Bags and mosaics: Semi-automatic identification of auxiliary verbal constructions for agglutinative languages. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 111–116.

Emese K. Molnár and Andrea Dömötör. 2023. Experiments on error detection in morphological annotation. In *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 186–190, Poznan. Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza.

Ahuvia Kahane and Martin Mueller. 2001. *The Chicago Homer*. University of Chicago Press/Northwestern University Library.

Ronald W. Langacker. 1987. *Foundations of cognitive grammar*. Stanford University Press, Stanford.

John Lee. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479, Prague, Czech Republic. Association for Computational Linguistics.

Lawrence Lessig. 2008. *Remix: Making Art and Commerce Thrive in a Hybrid Economy*. Penguin Press.

David Rosson, Eetu Mäkelä, Ville Vaara, Ananth Mahadevan, Yann Ryan, and Mikko Tolonen. 2023. Reception reader: Exploring text reuse in early modern british publications. *Journal of Open Humanities Data*, 9(1).

David A. Smith, Ryan Cordell, and Elizabeth Maddock Dillon. 2013. Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *2013 IEEE International Conference on Big Data*, pages 86–94.

Botond Szemes, Tímea Bajzát, Zsófia Fellegi, Péter Kundráth, Péter Horváth, Balázs Indig, Anna Diόssy, Fanni Hegedüs, Natali Pantyelejev, Sarolta Sziráki, Bence Vida, Balázs Kalmár, and Palkó Gábor. 2022. Az elte drámakorpuszának létrehozása és lehetőségei. In *Valós térben – Az online térért: Networkshop 31: országos konferencia*, pages 170–178, Budapest. HUNGARNET Egyesület.

Aleksi Vesanto, Filip Ginter, Hannu Salmi, Asko Nivala, and Tapio Salakoski. 2017. A system for identifying and exploring text repetition in large historical document corpora. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 330–333, Gothenburg, Sweden.

Three Studies on Predicting Word Concreteness with Embedding Vectors

Michael Flor
Educational Testing Service
Princeton, NJ, USA
mflor@ets.org

Abstract
Human-assigned concreteness ratings for words are commonly used in psycholinguistic and computational linguistic studies. Previous research has shown that such ratings can be modeled and extrapolated by using dense word-embedding representations. However, due to rater disagreement, considerable amounts of human ratings in published datasets are not reliable. We investigate how such unreliable data influences modeling of concreteness with word embeddings. Study 1 compares fourteen embedding models over three datasets of concreteness ratings, showing that most models achieve high correlations with human ratings, and exhibit low error rates on predictions. Study 2 investigates how exclusion of the less reliable ratings influences the modeling results. It indicates that improved results can be achieved when data is cleaned. Study 3 adds additional conditions over those of study 2 and indicates that the improved results hold only for the cleaned data, and that in the general case removing the less reliable data points is not useful.

Keywords: word concreteness, word embeddings, data reduction

1. Introduction

The importance of distinction between concrete and abstract concepts has been long noted in psycholinguistics (Paivio, Yuille, & Madigan, 1968). The so called 'concreteness effect' often finds that human participants process concrete words faster and more accurately than abstract words, in a variety of tasks, such as word naming, recognition, and recall, as well as sentence comprehension (Paivio 1991). Jessen et al. (2000) conducted fMRI studies indicating that concrete nouns are processed differently in the brain than abstract nouns.

Notions of concreteness and abstractness have also been used in computational approaches, both to investigate lexical relations, and for analysis of text. Concreteness of words has been widely used for metaphor detection (Maudslay et al., 2020; Köper and Schulte im Walde, 2017; Beigman Klebanov et al., 2015; Tsvetkov et al., 2014; Turney et al., 2011). For example, when a sentence describes an abstract agent performing a concrete action, it can be a strong indication of metaphorical usage. Choi and Downie (2019) used word concreteness scores to analyze trends in popular song lyrics across several decades, finding that concreteness in songs has been decreasing before the year 1991 and began increasing since then. Hills and Adelman (2015) analyzed distributions of word concreteness in published books; they noted "*a systematic rise in concrete language in American English over the last 200 years.*" Flor and Somasundaran (2019) investigated word concreteness in narrative writing of students, finding that concreteness positively correlates with rater scores of narrative quality.

Hill et al. (2014) analyzed the associations that concrete and abstract words have in a large corpus. They found that the more concrete words have smaller sets of context words, while abstract words have larger sets of context words. Naumann et al. (2018) investigated the concreteness of the contexts of concrete and abstract English words. They found that abstract words mainly co-occur with abstract words, but for concrete words cooccurrence patterns differ by part-of-speech. Tater et al. (2022) investigated selectional preferences of English nouns and verbs, and found that strong preferences exist with respect to concreteness and abstractness of subject and direct object slot fillers for verbs.

Early work in psycholinguistics has shown that concreteness/abstractness is not dichotomous but a matter of degree, and researchers began collecting human-assigned ratings for various words and producing lexical norms (Paivio et al., 1968). Presently three large human-rated datasets of concreteness are available for English (Coltheart, 1981; Brysbaert et al., 2014; Scott et al., 2019).

In parallel with utilizing the experimental ratings, researchers have also been interested in extrapolation of concreteness ratings to other words, for which ratings are yet unavailable. Notably, the interest in using computational linguistic approaches to extrapolate human semantic judgments is not limited to concreteness ratings, Methods to extrapolate ratings for a variety of variables, such as sentiment, arousal, and dominance, have been studied (Bestgen & Vincze, 2012; Turney & Littman, 2003); for a synthesis of some approaches see Mandera et al (2015).

Many researchers have reported that utilizing dense word representations (word embeddings) from distributional semantic language models can be useful for predicting and extrapolating concreteness values. Mandera et al. (2015) used several approaches to learn to predict psycholinguistic values from corpus data. For prediction of concreteness ratings, they used the data from Brysbaert et al. (2014). Using Random Forest learning over word vectors, they achieved a correlation of .781 with original scores, and even a higher correlation of .796 when using a KNN approach. Hollis et al. used regression over word2vec vectors and achieved a correlation r=.833. Paetzold and Specia (2016) used bootstrapped regression over word2vec embedding vectors from a corpus of 7 billion words. For predicting concreteness, their best result had Pearson

The Workshop on Cognitive Aspects of the Lexicon, pages 140–150
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

correlation of r=.862, with human ratings. Thompson and Lupyan (2018) used multiple linear regression over word vectors and obtained correlation of r=.86 with human ratings. Ljubešić et al. (2018) utilized word embedding vectors trained on Wikipedia to predict concreteness scores with SVM regression; they reported Spearman correlation of ρ=.872 between estimated and original values.

While human ratings provide the core data for extrapolation studies, such ratings are not without problems themselves. The published ratings for each word are usually average values across several human participants, and humans often disagree in their judgments; the standard deviations of human ratings per word vary considerably. Pollock (2018) provided an in-depth critique of crowd-sourced ratings of semantic psycholinguistic variables, such as concreteness, imageability, and emotional valence. Munoz-Rubke et al. (2018) have argued against using Likert-type rating scales for rating studies such as concreteness. Computational linguists have also noted problems with words for which human ratings show high disagreement (Tater et al., 2022; Beigman Klebanov et al., 2015).

In this paper we set to investigate to what extent words that have considerable rating disagreements influence word-embedding-based modeling of concreteness ratings. The paper is structured as follows. First, we describe the three large, published datasets of word-concreteness ratings for English. In study 1 we compare twelve word-embedding models as to their ability to model the concreteness ratings in those datasets. To the best of our knowledge this is the largest such comparison to date. In study 2, we pick two models and investigate how their predictions are influenced by exclusion of words with high standard deviations of concreteness ratings. In study 3 we introduce additional conditions on exclusion of such words, which shed light on their influence in the modeling process.

2. Datasets

The MRC Psycholinguistic Database (Coltheart, 1981; Wilson 1988) is one of the earliest large compilations of linguistic and psycholinguistic values for English words. It has concreteness ratings for 4295 English words, which were derived from experimentally established sets where participants rated words for perceptual concreteness on a 1-7 rating scale. In the MRC database they are expressed on a 100-700 scale (and rescaled back to 1-7 for the current study). Notably, the MRC database does not list the per-word standard deviations of the ratings.

Brysbaert et al. (2014) published a collection of 37,057 English words (mostly lemmas) with human-provided concreteness ratings (the BWK dataset). It is the largest such collection of ratings for English. The authors noted that previous collections of human concreteness ratings tended to focus too much on visual perception, and so for their rating study they emphasized all types of experiences (not only sensory, but also actions/activities). In that study, participants (native English speakers) received word lists and had to rate each word for concreteness, on a 5-point Likert scale, where only integer values could be chosen. After careful validation and filtering, the authors retained only those words that were known by at least 85% of the raters, and each word was rated by about 25 participants. The resulting concreteness score for each word is an average of the scores it received from its raters. The authors also released the standard deviation values of the ratings for each word. The BWK and MRC sets have an overlap of 3,935 words, and Pearson correlation of concreteness ratings is r=.919, a very high level of agreement.

Scott et al. (2019) published normative ratings for 5,553 English words on nine psycholinguistic dimensions, including concreteness. The authors called this data the Glasgow Norms (hereafter the GN dataset). In that study, for any given subset of words, the same participants provided ratings across all nine dimensions, and on average each word was rated by 33 participants. For concreteness, integer ratings were assigned on a 7-point Likert scale. Average concreteness values and standard deviations of ratings for each word were released by the authors. Some of the words in that study were polysemous and were presented with a disambiguator, e.g., *blubber (cry)* and *blubber (fat)*. By excluding the 871 such entries in that data, we utilize the 4682 single words (lemmas) that have concreteness ratings. The GN dataset has an overlap of 4,445 words with the BWK dataset, and Pearson correlation of average concreteness ratings between the sets is r=.93, indicating very high agreement of ratings.

For all three datasets, the published concreteness scores are real numbers, in the respective scale ranges. It is interesting to note the distribution of concreteness scores in the three datasets. Although the scales are of different magnitudes, it can be seen in the binned distributions (Figure 1) that the BWK data is skewed towards the more abstract side, while the GN data is skewed to the concrete side of the scale. The MRC has more words on the concrete side, but the extreme bins are 'underpopulated', especially the bin for very abstract words with scores in the range of 1-2.

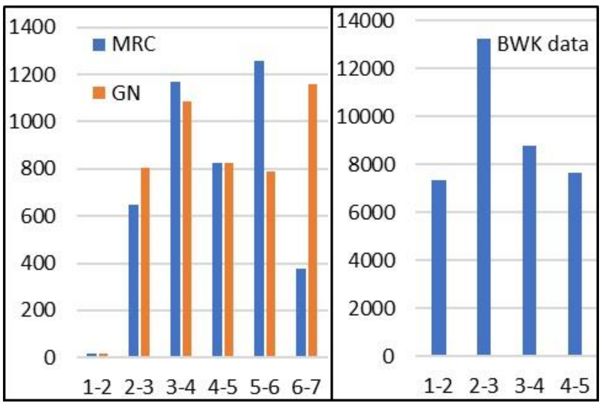


Figure 1: Binned distributions of concreteness scores in three datasets: MRC, GN, and BWK. Score bins on X-axes, word counts on Y-axes.

3. Experiments

3.1 Study 1

In Study 1 we investigate to what extent vector representations of words can be utilized for predicting word concreteness scores. Following previous studies (Thompson and Lupyan, 2018), we employ multiple linear regression as the learning method for the experiments. In such setting, the embedding vector dimensions serve as predictor variables.

We experimented with fourteen different embeddings models, as listed in Table 1. We included the widely used word2vec cbow model trained on Google News (Mikolov et al., 2013) and refer to it as *mikolov.w2v*. From Baroni et al. (2014) we adopted the word2vec cbow model trained on a window of 2 words (*baroni.w2v*), and also a vector model based on SVD of PMI word co-occurrence values (*baroni.pmi*). Two GloVe embeddings models (Pennington et al., 2014) were used – one trained on a corpus of 6 billion words (*glove.6b*) and a larger model, trained on 42 billion words (*glove.42b*). From the work of Levi and Goldberg (2014) we used two word2vec models trained on English Wikipedia data: one used a window of 5 words around a target word (*l&g.w5*), the other model used dependency parse relations (*l&g.deprel*). The *eigenwords* embeddings come from the work of Dhillon et al. (2015). *Lexsub* embeddings are a model introduced by Melamud (2015). The *ftWiki* model is a model trained on English Wikipedia, part of the Fast Text family of models (Bojanowski et al., 2017). The *paragram* model (Wietig et al., 2015) used a large database of English paraphrases to tune the word embeddings. Numberbatch embeddings (*nb17*) is a model based on the ConceptNet project data (Speer and Lowry-Duda, 2017). Two additional models use embeddings derived from Transformer architectures. We used the popular SentenceBERT library (Reimers and Gurevych, 2019) as an embedder, to produce static embeddings for the words in our experiments. The *MiniLM-L6-v2* model produces vectors of dimension 368, based on the BERT transformer model. The *distilroberta-v1* model produces vectors of dimension 768, derived from the DistilRoBERTa transformer model. In all experiments in this study, all vectors were normalized with L2 normalization.

It is worth noting that different vector models have different coverage for the words in the datasets (see Table 1). For the BWK data, among the classic models, the lowest coverage is by the *l&g.deprel* model, only 26,605 words (72% of the dataset), and the highest is by *glove.42b*, 35,491 words (96% coverage). Embeddings derived from SentenceBERT achieve full coverage, as such modern models can provide embeddings for any string. For the smaller MRC and GN datasets, the coverage was much better. Lowest coverage for MRC data was 4,140 words (96%), and for GN data: 4629 words (99%).

Experiments were performed for the MRC, BWK and GN datasets separately. All experiments involved 10-fold cross validation, with a 9:1 training:testing ratio. We used value clamping to prevent regression-based predicted values from falling outside of the original scales. Predicted values below 1 were reset to 1, and those above maxima (5 or 7) were reset to the max value.

Two evaluation measures were used to estimate the success of various models. One measure was Pearson correlation between the original published concreteness values and the predicted values. The higher the correlation, the better is the prediction. The other measure is Root Mean Square Error (RMSE), which measures the average squared difference between original and predicted scores. Lower values of RMSE indicate better prediction performance. Results (micro-averages) for all the experiments are presented in Table 1. A single model, *nb17*, achieved the best results in all datasets, on both the correlation and the RMSE measures.

Results for the BWK dataset indicate that all models show rather impressive prediction power – correlations ranging above 0.8 (except *glove.6b*), but none reaches 0.9. Across different language models, the RMSE values for BWK data range between 0.472 and 0.633. Divided by the scale range, 4, those RSMEs are at a magnitude of 12-16% of the score range.

Results for the GN dataset indicate that all models show very strong results, all correlations range above 0.8, and two models – *nb17* and *lexsub* achieve correlations above 0.9. The RMSE values for the GN data range from 0.572 to 0.872. Those values are larger than values obtained for the BWK data. However, GN data was rated on a 1-7 scale, and so higher error values should be expected. If we divide RMSE values by the scale range, we can see that the error results in the two experiments are comparable. Lowest RMSE values: for BWK data 0.472/4=0.118; for GN data 0.572/6=0.095. The highest RMSE: for BWK: 0.633/4=0.158; for GN: 0.872/6=0.145.

The results for MRC data resemble those of GN data, although each language model achieves slightly worse (lower) Pearson correlation values for MRC than for GN, but slightly better (lower) RMSE values for MRC than for GN data.

3.2 Study 2

The background for Study 2 stems from the criticism that some researchers have pointed toward the reliability of psycholinguistic ratings with Likert-type scales. Munoz-Rubke et al. (2018) have noted that when participant ratings are averaged and assigned as final word scores, for categories such as concreteness, the approach may have important limitations, as the results can be highly distorted by outliers. Specifically for concreteness values norms from the Brysbaert et al. (2014) study, Pollock (2018) has argued that the mean concreteness values for words do not reflect the judgments that actual participants made: "*this problem applies to nearly every word in the middle of the concreteness scale.*"

		BWK data			GN data			MRC data		
Model name	**dims**	**coverage**	**Pearson**	**RMSE**	**Coverage**	**Pearson**	**RMSE**	**coverage**	**Pearson**	**RMSE**
sbert Mini-LM6-v2	368	37057	0.825	*0.573*	4681	0.858	*0.736*	4295	0.812	*0.708*
sbert distilroberta-v1	768	37057	0.815	*0.588*	4681	0.797	*0.872*	4295	0.752	*0.807*
mikolov.w2v	300	33975	0.848	*0.539*	4629	0.887	*0.661*	4220	0.854	*0.627*
nb17	300	35488	**0.885**	***0.472***	4679	**0.917**	***0.572***	4292	**0.883**	***0.569***
glove.42b	300	35491	0.821	*0.579*	4682	0.855	*0.745*	4290	0.814	*0.704*
glove.6b	300	31619	0.783	*0.633*	4680	0.825	*0.811*	4261	0.784	*0.752*
l&g.deprel	300	26605	0.868	*0.507*	4651	0.891	*0.649*	4195	0.874	*0.588*
ftWiki	300	35319	0.850	*0.535*	4680	0.878	*0.686*	4294	0.849	*0.640*
Lexsub	600	28274	0.874	*0.496*	4677	0.905	*0.612*	4232	0.879	*0.577*
Eigenwords	200	28276	0.865	*0.511*	4651	0.883	*0.673*	4140	0.867	*0.601*
l&g.w5	300	27212	0.834	*0.563*	4657	0.870	*0.706*	4214	0.850	*0.636*
Paragram	300	35308	0.805	*0.601*	4682	0.811	*0.840*	4286	0.773	*0.768*
baroni.ppmi	500	30260	0.839	*0.555*	4681	0.877	*0.690*	4261	0.851	*0.638*
baroni.w2	400	30260	0.811	*0.596*	4681	0.880	*0.688*	4261	0.841	*0.656*

Table 2: Results of word-concreteness score prediction for three datasets, with 14 different vector-space models. *Dims* is the number of dimensions per vector. The columns labeled *Coverage* are counts of words that had vectors in the respective language model. For Pearson correlations, higher value means better prediction; for RMSE, lower value means better prediction.

He recommended that researchers who use such ratings pay attention to the standard deviations of ratings and use only the stimuli for which standard deviations are as low as possible. The relevance of such critique to our work is quite direct. What would happen if we excluded from our data all items (words) that are 'less reliable'? Would it improve the concreteness prediction models? On the other hand, excluding some data would make the datasets smaller; and having less data may lead to inferior learning.

Beigman Klebanov and Beigman (2014) and Jamison and Gurevych (2015) have suggested that, in supervised machine learning, the presence of difficult items in the training sets is detrimental to learning performance and that performance can be improved if systems are trained on only easy data. They define ‘easy’ as less controversial in human annotations. This seems exactly analogous to our current case. Words that have high standard deviations (SD) of human-rated concreteness are 'less reliable' as to their real concreteness value, they are more 'difficult' cases. Excluding them from the training data may leave just the more reliable, 'easier' data for learning and thus might lead to improved model performance.

Standard deviations of rating values for each word are available for the BWK and the Glasgow Norms datasets. To understand the potential scope of data reduction, we plot the number of words in each dataset as a function of different SD value thresholds, and also by score-bins of the ratings. Figure 2 (left panel) presents the plot for the BWK dataset. The black bars represent the data when nothing is excluded, corresponding to Figure 1. The red bars indicate the counts of remaining words when all words with SD>1.5 are excluded. Such exclusion affects mostly words in the score bins 2-3 and 3-4. The green bars indicate the counts when all words with SD>1.2 are excluded. Again, we can see that the largest data reduction occurs for words in the score bins 2-3 and 3-4. With exclusion threshold of SD>1.0 (maroon-color bars), almost all words in bins 2-3 and 3-4 get excluded. The exclusion rates are much more gradual for score bins 1-2 and 4-5, which are closer to the extremes of the concreteness rating scale.

Figure 2 (right) presents the distributions of words for the GN dataset. The black bars represent the data when nothing is excluded, corresponding to Figure 1. Data reduction thresholds for this set are somewhat different. At the exclusion threshold of SD>1.0, only the bin of scores 6-7 retains some considerable number of words, while all other bins are almost emptied. Data reduction is especially dramatic for bins of scores 3-4 and 4-5 (the middle of the rating scale).

The design for study 2 is as follows. We investigate how gradual elimination of some data from the datasets influences the quality of the learned models. For each dataset, we exclude all words that exceed a given SD threshold and train a multiple regression model with 10-fold cross-validation. This mode of exclusion is systematic. For the sake of comparison, we also check what happens if the same number of words are excluded, but chosen randomly, rather than by an SD threshold. For example, for BWK data (37,057 words), for a threshold of SD≤1.4 we exclude 7,053 words, and experiment (full 10-fold cross-

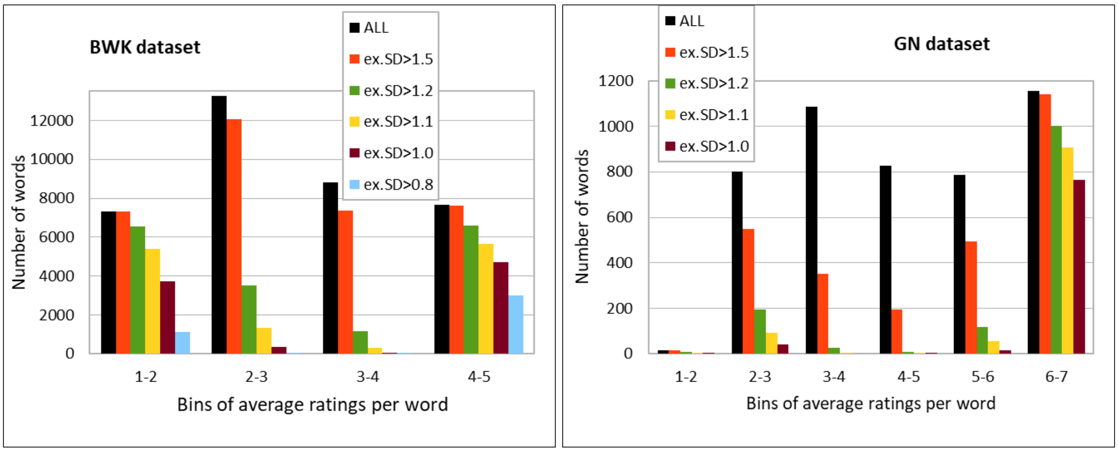


Figure 2: Binned distributions of concreteness scores, data exclusion by SD thresholds.

validation) with the remaining 30,004 words. In a matching control condition, we exclude 7,053 words randomly chosen, and run the experiment with the remaining 30K words.

For the BWK dataset, the systematic exclusion steps are from SD value of 1.0 to 0.6 with a step of 0.1 (more data is excluded on each step). For the GN data, the SD thresholds are from 2.0 to 1.0 with a step of 0.1. For each dataset we also use a condition where no words are excluded, as in study 1. Table 2 presents the counts of remaining words for each condition.

Inclusion	BWK data	GN data
All data	37057 (100%)	4682 (100%)
SD≤2.0		4599 (98%)
SD≤1.9		4485 (96%)
SD≤1.8		4260 (91%)
SD≤1.7	36942 (99%)	3901 (83%)
SD≤1.6	36345 (98%)	3383 (72%)
SD≤1.5	34375 (93%)	2748 (59%)
SD≤1.4	30004 (81%)	2198 (47%)
SD≤1.3	23916 (64%)	1738 (37%)
SD≤1.2	17814 (48%)	1357 (29%)
SD≤1.1	12681 (34%)	1069 (23%)
SD≤1.0	8847 (24%)	825 (18%)
SD≤0.9	6118 (17%)	
SD≤0.8	4147 (11%)	
SD≤0.7	2860 (7%)	
SD≤0.6	1998 (5%)	

Table 2: Number of remaining words in two datasets, by inclusion thresholds on SD values.

For Study 2 we use two embedding models from Study 1 that have good performance and also have good lexical coverage over the BWK and GN datasets – *nb17* and *sbert MiniLM-L6-v2*. Note that the number of words used in each experimental condition, as presented in Table 2, applies only to the *sbert* model, as it has full coverage of the datasets; *nb17* has lower coverage and thus the number of words used is slightly lower in each respective condition. Just as in study 1, RMSE and Pearson correlation are used as evaluation measures in study 2.

Results for the BWK dataset are presented in Figure 3. The correlation results with *sbert* and *nb17* are quite similar (Figure 3, left panels). When very little data is excluded (thresholds 1.7 and 1.6), the results of systematic or random exclusion are quite the same, and very close to those of no exclusion. However, the results begin to separate from threshold 1.5. The results from systematic exclusion become higher and higher with each successive exclusion threshold, they reach beyond correlation of .9, and for *nb17* – even beyond .95. The peak results are achieved at SD≤0.8. After that threshold, the correlation values begin decreasing, though they are still higher than for the full dataset. For the control conditions with random exclusion, the correlation values do not improve with successive exclusions, they even have a slight tendency of decreasing, and never get higher than values for the full-data condition.

RMSE results for the BWK dataset are presented in Figure 3, right-side panels. Note that for RMSE, lower error values indicate better performance. The results with *nb17* and *sbert* are quite similar. For inclusion thresholds 1.7 to 1.3, the RMSE results for systematic or random exclusion are very close to each other, and approximately the same as under the no-exclusion condition. However, as more and more data gets excluded, RMSE values for systematic exclusion begin decreasing; the decrease even accelerates (the black-color lines curve down), whereas the error levels for random exclusion (orange-colored lines) remain the same, or even increase slightly. Notably the separation of results between systematic and random conditions begins at SD≤1.2 for the *nb17* model and at SD≤1.0 for the *sbert* model.

Results for the GN dataset are presented in Figure 3. The correlation results with *nb17* and *sbert* are quite similar (Figure 3, left panels). The trends are also similar to those of the BWK dataset results. When very little data is excluded (thresholds 2.0 to 1.8), the results of systematic or random exclusion are quite similar, and very close to those of the no-exclusion condition. For further thresholds, systematic exclusion leads to higher correlation results, until threshold levels of 1.3 or 1.2, and then the correlation results start decreasing quite sharply. The sharp decreases might be due to the sharp reduction in the size of the dataset, or due to the dramatic change in the distribution of values in the reduced corpus (see Figure 2). The best correlation results are, for *sbert*: r=.887; for *nb17*: r=.944; all when SD≤1.4. Under the random exclusion, the correlations tend to decrease.

The RMSE results for the GN data are presented in Figure 3, right-side panels. The results for systematic exclusion are similar to the RMSE results in the BWK dataset – at first the error levels are quite similar to those under the no-exclusion condition, but then the RMSE values get increasingly lower and lower (black-color lines tend to curve down). The results for random exclusion are markedly different from systematic exclusion. At first the error levels are close to those under the no-exclusion condition, but then the RMSE values begin rapidly increasing (orange lines curve up), indicating worsening performance.

The results of study 2 indicate that when the 'less reliable' data is excluded from the datasets, regression models based on word-embeddings can achieve much better results that with the full data.

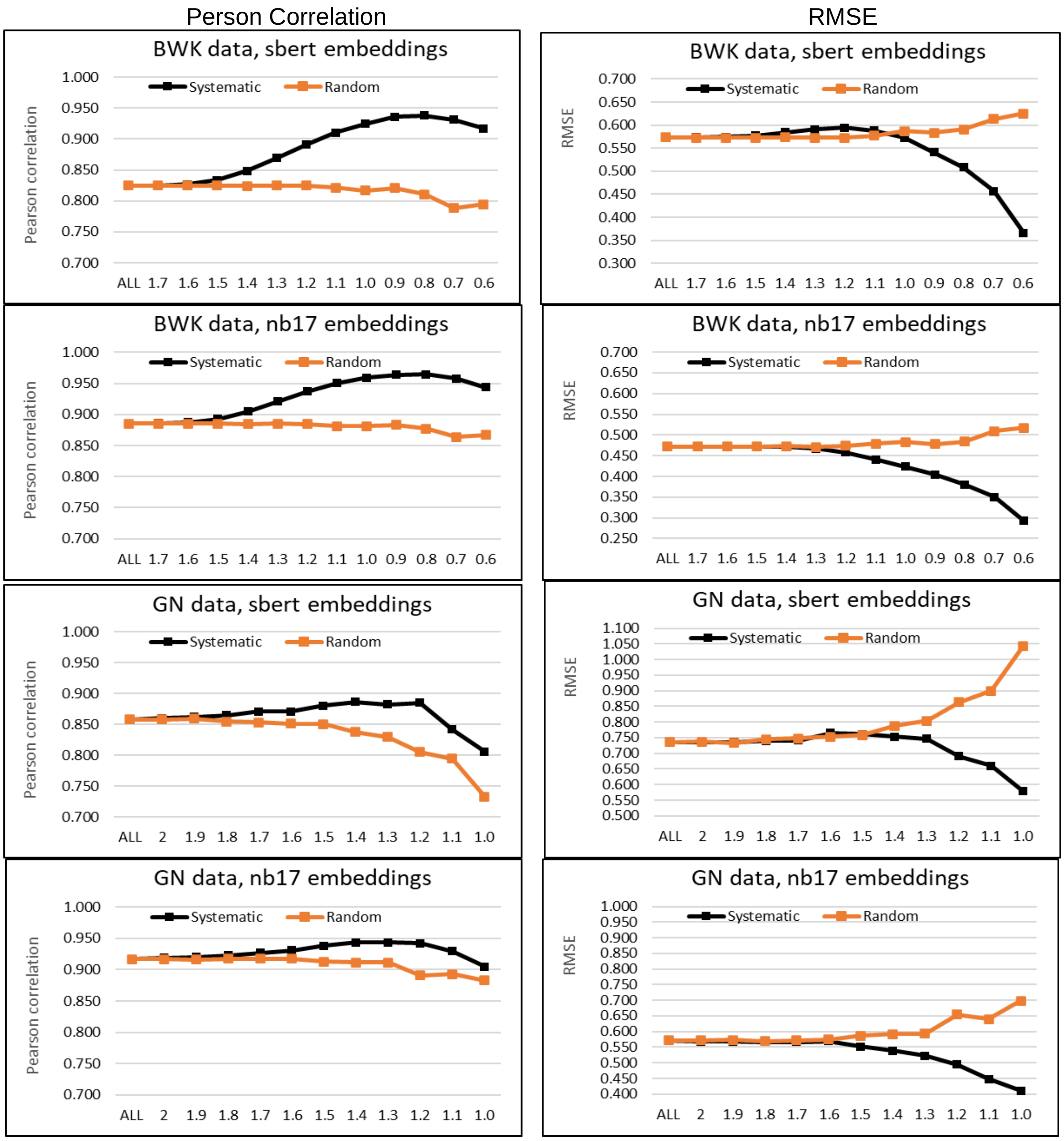


Figure 3: Pearson correlations (left) and RMSE (right) for predicting concreteness scores on two datasets, as a function of data reduction (systematic by SD thresholds, or random), using two different language models. Data points marked 'ALL' represent a condition where all available words were included.

3.3 Study 3

An important aspect in study 2 was the change in distribution of concreteness values, under the systematic exclusion condition. Do models achieve better results because they learn on increasingly 'cleaner' data, or simply because of the different distribution of values? And what about the 'less reliable' values? Should we exclude them from modeling at all? How would models trained on cleaner, reduced data perform on unfiltered data? Study 3 addresses those questions.

In study2, we tested what happens when the dataset was successively reduced. Under systematic reduction, both the training folds and the testing folds were reduced as per the SD thresholds. In the other condition, random exclusion was used for all folds. In study 3 we add two new conditions where we mix the data exclusion methods. In a condition called SR, data reduction in training folds uses systematic reduction (by SD thresholds), but in the testing folds a comparable amount of data is excluded randomly. This condition evaluates what happens when training data is systematically cleaned (and the distribution of concreteness scores changes), but the testing data is just randomly reduced, and so it keeps the same distribution as the whole dataset. Under another condition, called RS, we reverse the reduction methods. Data for the testing folds is reduced systematically (by SD thresholds), but the training folds get a proportional random reduction. Thus, the models are trained on approximately the same distribution as the whole dataset, but are tested on just the 'cleaner' data. The overall amounts of included data decrease in the same way under the new conditions, just as in study 2. For study 3 we used the same datasets and same vectors as in study 2. All experiments were run with 10-fold cross-validation. Results are presented in Figure 4. For ease of comparison, the results from study 2 are shown again, with the results of the new conditions added (yellow lines for SR and green lines for RS).

When models are trained on increasingly 'cleaner' data, their ability to predict values for 'non-cleaned' data (yellow lines) keeps up with models that do not 'clean' the data (red lines), both for correlation and RMSE. However, after certain levels of data reduction the 'clean'-trained models begin losing it – they achieve slightly lower correlations and make dramatically larger errors, as compared to models that train and test on randomly-reduced data (red lines). Comparing yellow lines to black lines (in both cases models train on cleaned data) shows that the composition of the test data makes a huge difference – when test data is also clean, the best overall results are achieved, but when the test data is unfiltered, the worst results are achieved (lowest correlations and largest errors).

Next, we consider models that test on just the clean data, but train on cleaned (black) or unfiltered data (green lines). Looking left to right on each panel (left side) in Figure 4, the green line keeps up with the black line until SD 1.2 (BWK) or SD 1.4 (GN), reduction to about 47% of the full data. After that the green lines show worse results than the black lines, but still better than the other lines. It seems that the models trained on unfiltered data retain most of the information needed to predict clean data; that is until the distributions become so different that prediction deteriorates (the respective RMSE values start rising while correlations get lower).

4. Discussion

Many previous studies used the large BWK dataset. Thompson and Lupyan (2018) reported a correlation of r=.86; Hollis et al. (2017) reported a correlation of r=.833; Mandera et al. (2015) obtained a correlation of r=.781. Ljubešić et al. (2018) reported Spearman correlation ρ=.887 on BWK data and ρ=.872 on MRC data. Paetzold and Specia (2016) reported a correlation of r=.862 on MRC data. Our results in study 1 indicate that comparable or better prediction levels can be obtained with several different language models, using ordinary multiple regression. While previous studies have used BWK and MRC datasets, the current study is first to also use the Glasgow Norms data for concreteness prediction. The results resemble those of BWK and MRC data. None of the previous studies used RMSE as an evaluation measure for concreteness ratings prediction. In study 1, RMSE results for BWK data are typically lower than for MRC and GN data, probably due to differences in scales. Beyond that, RMSE results for different embedding models are quite similar to correlation results – embeddings that get better correlations also show lower error results.

Study 2 was motivated by the notion of unreliable word concreteness ratings, which reflect considerable disagreements among human raters. In the BWK dataset, less than 50% of the words have standard-deviation values below 1.2, and only 24% have SD values below 1.0. In the GN dataset less than 29% of the words have SD values below 1.2 and just 18% have values below SD 1.0. Study 2 investigated how exclusion of unreliable data points influences regression modeling. It was found that when human raters agree more on concreteness of words, such ratings can be modeled/predicted very well with vector space models. Higher correlations and lower errors are obtained as compared to learning on the full data.

However, the distributions of concreteness scores in the BWK and the GN datasets change drastically when less reliable words are excluded – most unreliable words are in the middle of the distributions and are excluded with successive data cleaning. Study 3 investigated whether results in study 2 were due to changes in concreteness score distributions. The results showed that training on clean data does not generalize well to unfiltered data, especially with regard to magnitude of errors (RMSE). On the other hand, training on unfiltered data and testing on just the clean data reveals that the models have enough information to predict scores for clean data, especially

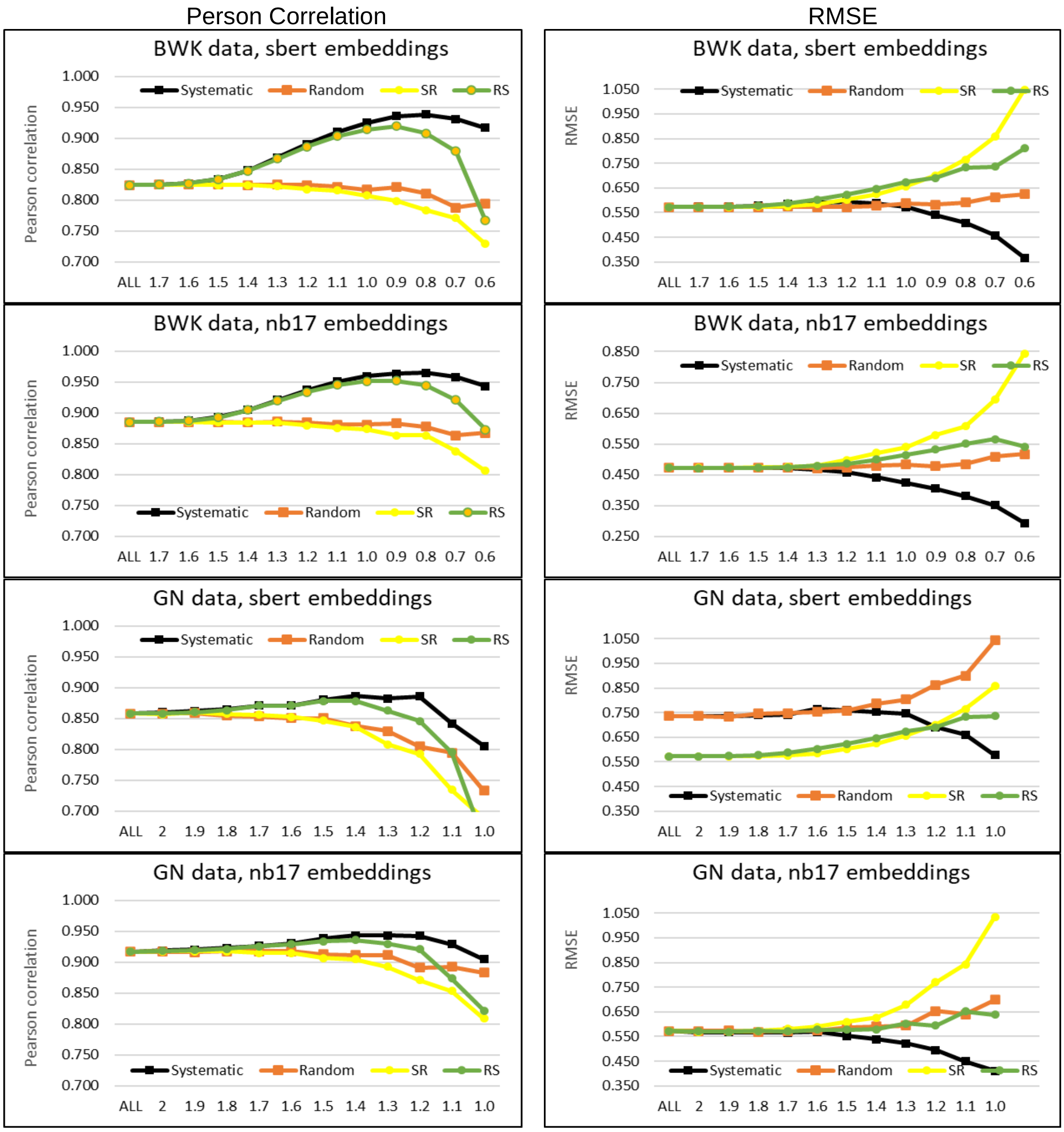


Figure 4: Pearson correlations (left) and RMSE (right) for predicting concreteness scores on two datasets, with four methods of data reduction, using two different language models.

on the correlation measure, but less so on RMSE. Thus, if we are only interested in predicting concreteness for 'reliable' words, cleaning the training data can be useful. If the potential 'reliability' of new words is unknown (as would be the case for most new words), filtering the training data is not recommended.

We continue the discussion in relation to the distinction between easy and difficult cases. Uma et al (2021) provide an extensive review on the influence of hard cases for machine learning, where difficulty arises from disagreements in human annotation. They provide a taxonomy of potential reasons for disagreement. Among the sources of disagreement, Uma et al. mention a) annotator/rater errors b) problems with the task interface, c) problems with task definition, d) situational item difficulty, e) genuine ambiguity of the data, and f) rater subjectivity.

Annotator/rater errors can be mistakes or slips made due to inattention, or other random factors. Interface issues can arise when task interface may have technical complications (e.g., selecting text spans). Problems with task definition may lead to disagreements when the task is not well defined, includes vague statements, or, in case of classification, classes that are not mutually exclusive. Item difficulty (for rating/annotation) relates to cases when the interpretation of the data is unclear. For example, for image labelling, if the image is too blurred, annotators may disagree as to what they actually see there, and thus disagree on a label. In the task of textual entailment, an item may be difficult because the text is convoluted and a core assertion is not easy to discern. As noted by Uma et al., the problem is not that an item lacks a 'true' label, only that the 'true' label is difficult to distinguish. As they note, the conclusion either follows from the premise, or it doesn't, but not both. This contrasts with the ambiguity category, where the data items can be truly ambiguous, i.e., have different valid interpretations. Uma et al. mention that ambiguity cases have been shown to arise in annotation of anaphora and of POS tags. The final category, subjectivity of judgement, relates to cases where annotators/raters hold different opinions. The prototypical example is annotation of offensive language, where annotators may disagree on whether a given expression is offensive, and different opinions can be simultaneously valid.

Notably, the above taxonomy was developed in relation to disagreements on tasks that involve data classification, and the labels are on nominal scales (but the amount of disagreement can be expressed on continuous scales). Uma et al. (2021) presented several studies around the question on how to integrate disagreements into machine learning processes. There were few studies with data on other scales. The study by Jamison and Gurevych (2015) included a dataset on biased language, where the labels were on an ordinal scale (*no bias, some bias, very biased*), and a dataset on affect recognition for text snippets, with a scoring scale of 0-100. Loukina et al. (2018) investigated automated speech scoring (for language proficiency assessment), where spoken segments were scored on a 1-4 integer scale. In both studies the question was whether training on the easier data (with clear-cut cases and less disagreement) would be beneficial for training ML systems. The results were mixed. Jamison and Gurevych found that for data on nominal scales (classification tasks), training on easier data leads to improved performance. For affect data, training on easier cases can lead to improved results, when testing on easy cases, but only marginal or no improvement when testing on all data or just the hard cases. Loukina et al. found that training on easy data (as compared to mixed data) did not lead to better performance on test data. On the other hand, they found that the choice of data for testing the systems did matter – performance on easier testing data was always better than performance of mixed testing data. Yet, evaluation on just the easier cases should not be dismissed, as it provides an important validity indicator: making many errors on difficult cases might be tolerable, making many errors on clear-cut cases may raise serious doubts about validity of the system.

It is interesting to note how ratings of psycholinguistic variables, such as concreteness, valence, affect, etc., relate to the above taxonomy of rater disagreements. Concreteness scores from human raters are typically obtained on Likert scales. While attention and other random errors might be involved, Munoz-Rubke et al. (2018) also mention potential outlier effects. There could also be issues with reliability of raters (though responses from unreliable raters were eliminated in the Brysbaert et al. (2014) study). Task definition for rating concreteness/abstractness has also been criticized. Brysbaert et al. (2014) made special emphasis in rater instructions on concreteness in other modalities beyond visual perception, however, their results do not differ much from MRC and GN datasets, where such instructions were not explicitly presented. Attributing rater disagreement in concreteness ratings to 'situational item difficulty' is not quite plausible since ratings involved single words. A more plausible explanation for disagreement may be in the genuine ambiguity of some words, and/or the very subjective nature of concreteness ratings (Pollock, 2018).

Cases of ambiguity may arise when words have multiple senses or even just different parts of speech. For example, in the BWK dataset (scale 1-5), the word '*official'* has concreteness of 2.53 and SD of 1.43, while '*officially*' has concreteness 1.63 and much lower SD of 0.83. It might be that some raters interpreted '*official*' as a noun (and thus denoting a person), while others considered the adjective meaning (which is more abstract). The word '*officially*' is related to the same core meaning but has no such ambiguity. Perhaps concreteness ratings should be assigned per sense and not per wordform. Indeed, the Glasgow Norms (scale 1-7) have taken an early step in that direction, where 871 polysemous words were presented with a disambiguator, and thus the concreteness rating is per sense. However, even in such a disambiguated subset considerable variability of individual ratings exists – 360 entries on that list have SD>1.5, and the average SD of the disambiguated subset is 1.36. It seems raters disagreed even while rating specific senses of words.

The notion of collecting ratings per word sense is also related to predicting concreteness from word embeddings. Most of the classical word embeddings datasets (such as Google News word2vec, GloVe, etc) are not sense disambiguated, and their embeddings represent either a mix of senses or the most prevalent senses of words. For compatibility with such data, we opted to use Sentence-BERT embeddings in a similar way (i.e., per wordform). We opted to not use contextual BERT (or similar) embeddings per word and average them across multiple contexts. The issue in such case would be which contexts should be used for such averaging, and whether selection of contexts could have an influence on the senses that are implicitly modeled. However, this path that was not taken is also a path for future research. By carefully selecting contexts over which one averages contextual embeddings, a researcher might thus obtain sense-specific vectors, and potentially model sense-specific concreteness (and other psycholinguistic variables).

In sum, there seems more future work might be needed, both for collecting more reliable concreteness ratings, and for developing more sophisticated computational models of concreteness.

5. Conclusion

We investigated modeling of word-concreteness ratings with word embeddings. Study 1 demonstrated that human-produced concreteness scores can be successfully predicted by using ordinary multiple regression with word embeddings. We compared 14 embedding models over three different datasets of human-produced concreteness scores. In all cases we obtained high Pearson correlation values (between .8 and .9) between original and estimated ratings. Using the RMSE evaluation measure, we find that all models achieve relatively low average error levels (mostly ranging from .5 to .8), which translates to 10-15% on the corresponding rating scales. Studies 2 and 3 investigated the effect of words that have 'less reliable' human-ratings. Rater disagreements for any given word result in higher standard-deviation of scores for that word. Using two datasets where standard deviation values for each word were released, we investigated how exclusion of words with high standard deviation values affects embedding-based regression models that learn to estimate the concreteness scores for words. We find that systematic exclusion of 'less reliable' words from the learning data can lead to evident improvement of results. However, study 3 indicates that such improvements stem from drastic changes in the distribution of concreteness scores when data is 'cleaned'. Training on filtered data does not generalize well to unfiltered data, whereas training on unfiltered data has enough information for modeling values for clean data.

6. Acknowledgments

We thank Beata Beigman Klebanov and two anonymous reviewers for valuable comments that helped to improve this paper.

7. Bibliographical References

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the ACL* (Volume 1: Long Papers), pages 238–247, Baltimore, Maryland.

Beigman Klebanov, B., Leong, Ch.W., Flor, M. (2015). Supervised Word-Level Metaphor Detection: Experiments with Concreteness and Reweighting of Examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20.

Beigman Klebanov, B., and Beigman, E. 2014. Difficult Cases: From Data to Learning, and Back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Short Papers), pages 390–396.

Bestgen, Y., and Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, 44(4):998–1006.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (20170. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 64:904-911.

Choi, K., and Downie, J.S. (2019). A Trend Analysis on Concreteness of Popular Song Lyrics. In *Proceedings of the 6th International Conference on Digital Libraries for Musicology* (DLfM '19), pp. 43–52.

M. Coltheart (1981), The MRC Psycholinguistic Database, *Quarterly Journal of Experimental Psychology*, 33A:497-505.

Dhillon, P.S., Foster, D.P., Ungar, L.H. (2015). Eigenwords: Spectral Word Embeddings. *Journal of Machine Learning Research,* 16:3035-3078.

Flor, M., and Somasundaran, S. (2019). Lexical concreteness in narrative. In *Proceedings of the Second Storytelling Workshop*, pages 75–80. Florence, Italy, August 1, 2019.

Hill, F., Korhonen, A., and Bentz, C. (2014). A Quantitative Empirical Analysis of the Abstract/Concrete Distinction. *Cognitive Science*, 38(1):162–177.

Hills, T.T., and Adelman, J.S. (2015). Recent evolution of learnability in American English from 1800 to 2000. *Cognition*, 143:87–92.

Hollis, G., Westbury, C., and Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, 70(8):1603–1619.

Jamison, E.K. and Gurevych, I. (2015). Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297.

Jessen, F., Heun, R., Erb, M., Granath, D.O., Klose, U., Papassotiropoulos, A., and Grodd, W. (2000). The concreteness effect: Evidence for dual coding and context availability. *Brain and Language*, 74(1), 103–112.

Köper M., and Schulte im Walde, S. (2017). Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the First Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30.

Levy, O., and Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), pages 302–308, Baltimore, Maryland.

Ljubešić, N., Fišer, D., and Peti-Stantić, A. (2018). Predicting Concreteness and Imageability of Words Within and Across Languages via Word

Embeddings. In *Proceedings of the 3rd Workshop on Representation Learning for NLP*, pages 217–222. Melbourne, Australia.

Loukina, A., Zechner, K., Bruno, J., and Beigman Klebanov, B. (2018). Using exemplar responses for training and evaluating automated speech scoring systems. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–12, New Orleans, Louisiana.

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology*, 68(8):1623–1642.

Maudslay, R. H., Pimentel, T., Cotterell, R., and Teufel, S. (2020). Metaphor Detection Using Context and Concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 221–226.

Melamud, O., Dagan, I., and Goldberger, J. (2015). Modeling Word Meaning in Context with Substitute Vectors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 472–482, Denver, Colorado.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR (Workshop Poster)* 2013.

Munoz-Rubke, F., Kafadar, K., and James, K. H. (2018). A new statistical model for analyzing rating scale data pertaining to word meaning. *Psychological Research*, 82:787–805.

Naumann, D., Frassinelli, D., and Schulte im Walde, S. (2018). Quantitative Semantic Variation in the Contexts of Concrete and Abstract Words. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, pages 76–85, New Orleans, LA, USA.

Paetzold, G.H., and Specia, L. (2016). Inferring Psycholinguistic Properties of Words. In *Proceedings of NAACL-HLT 2016*, pages 435–440, San Diego, California.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3):255–287.

Paivio, A., Yuille, J.C. and Madigan, S.A. (1968). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology* Monograph Supplement, 76.

Pennington, J., Socher, R., and Manning, C.D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 1532–1543, Doha, Qatar.

Pollock, L. (2018). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods*, 50:1198–1216.

Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Scott, G.G., Keitel,A., Becirspahic, M., Yao, B., and Sereno S.C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51(3), 1258–1270.

Speer, R. and Lowry-Duda, J. (2017). ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada.

Tater, T., Frassinelli, D., and Schulte im Walde, S. (2022). Concreteness vs. Abstractness: A Selectional Preference Perspective. In *Proceedings of the AACL-IJCNLP 2022 Student Research Workshop*, pages 92–98.

Thompson, B, and Lupyan, G. (2018). Automatic Estimation of Lexical Concreteness in 77 Languages. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Tsvetkov Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 248–258, Baltimore, Maryland.

Turney, P.D., and Littman, M.L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Turney, P.D., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*, pages 680–690. Edinburgh, Scotland, UK.

Uma, A.N., Fornaciari, T., Hovy, D., Paun, S., Planck, B., and Poesio, M. (2021). Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72, 1385-1470

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). From Paraphrase Database to Compositional Paraphrase Model and Back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers,* 20:6–10.

Combining Neo-Structuralist and Cognitive Approaches to Semantics to Build Wordnets for Ancient Languages: Challenges and Perspectives

Erica Biagetti, Martina Giuliani, Silvia Zampetta, Silvia Luraghi, Chiara Zanchi
University of Pavia, Universities of Pavia and Bergamo, University of Pavia, University of Pavia, University of Pavia
erica.biagetti@unipv.it, martina.giuliani@unibg.it, silvia.zampetta01@universitadipavia.it, luraghi@unipv.it, chiara.zanchi@unipv.it

Abstract

This paper addresses challenges encountered in constructing lexical databases, specifically WordNets, for three ancient Indo-European languages: Ancient Greek, Latin, and Sanskrit. The difficulties partly arise from adapting concepts and methodologies designed for modern languages to the construction of lexical resources for ancient ones. A further significant challenge arises from the goal of creating WordNets that not only adhere to a neo-structuralist relational view of meaning but also integrate Cognitive Semantics notions, aiming for a more realistic representation of meaning. This integration is crucial for facilitating studies in diachronic semantics and lexicology, and representing meaning in such a nuanced manner becomes paramount when constructing language resources for theoretical research, rather than for applied tasks, as is the case with lexical resources for ancient languages. The paper delves into these challenges through a case study focused on the TEMPERATURE conceptual domain in the three languages. It outlines difficulties in distinguishing prototypical and non-prototypical senses, literal and non-literal ones, and, within non-literal meanings, between metaphorical and metonymic ones. Solutions adopted to address these challenges are presented, highlighting the necessity of achieving maximum granularity in meaning representation while maintaining a sustainable workflow for annotators.

Keywords: WordNet, ancient Indo-European languages, relational semantics, cognitive semantics, temperature domain

1. Introduction

In this paper, we delve into some challenges encountered while building three lexical databases, specifically WordNets, for three ancient Indo-European languages: Ancient Greek, Latin, and Sanskrit (Biagetti et al. 2021). These issues are partly related to adapting a set of concepts and methodologies designed for modern languages to constructing lexical resources for ancient ones, thus without relying on native speakers' support.

Crucially, another set of challenges stems from our programmatic goal of constructing WordNets that not only adhere to a neo-structuralist relational view of meaning (Geeraerts, 2010: 124-126, 158-160) but also integrate notions of Cognitive Semantics (e.g., Taylor, 2003; Aitchinson, 2003). This integration should allow for a more fine-grained and "more realistic" representation of meaning (Geeraerts, 2001: 18-19; 2007: 1168), thus facilitating studies in diachronic semantics and lexicology. In principle, a representation of meaning of this sort is of primary importance when constructing language resources that are not primarily aimed at applied tasks but rather at theoretical research, as are lexical resources of ancient languages. The paper discusses the latter set of issues through the lens of a case study, specifically examining the meanings associated with words pertaining to the TEMPERATURE conceptual domain in the three languages[1].

The paper is organized as follows. In Section 2 we present the new family of WordNets for ancient Indo-European languages. Specifically, in Section 2.1 we introduce the main features of WordNets, specifying the types of semantic information they contain and those they do not. In Section 2.2 we explain how we enhanced our WordNets with notions of Cognitive Semantics and present the potential of this approach. Section 3 contains a discussion of the challenges encountered in such implementation, taking the lexicon of TEMPERATURE as a case study. In particular, we outline the difficulties faced in distinguishing between prototypical and non-prototypical senses, literal and non-literal ones, and – among non-literal senses – between metaphorical and metonymic ones. We also present the solutions adopted to address these challenges. Section 4 contains the conclusions.

2. A family of WordNets of Ancient Indo-European Languages

2.1 WordNets: What Semantic Information They Contain, What They Do Not

WordNet is a lexical database that stores meaning in a network, initially designed by the psycholinguists George Miller and Christiane Fellbaum (Fellbaum, 1998; Miller and Fellbaum, 2007; Miller et al., 1990) and compiled for English at Princeton University. However, it soon lost its psycholinguistic flair and became a project in computational lexical semantics. Since the first Princeton WordNet, similar databases have been built (e.g., Vossen 1998, 2004) or are currently being built for many other languages, including ancient ones such as Latin, Ancient Greek, Sanskrit, and Old English. Researchers have further attempted to link these WordNets to larger language resource infrastructures (Biagetti et al. 2021; Khan et al., 2023; for Latin, see Bizzoni et al., 2014; Minozzi, 2017; Franzini et al., 2019; Mambrini et al., 2021; for Ancient Greek, see Boschetti, 2019; Zanchi et al., 2021; for Sanskrit, see Hellwig, 2017; Old English: Khan et al., 2022). Nowadays, the Global WordNet Association (available at http://globalwordnet.org/) promotes a collective forum for the standardization of existing WordNets, as well as for the development of shared guidelines and methodologies for building new WordNets and related linguistic resources.

[1] In Cognitive Semantic scholarship, concepts, conceptual domains and conceptual metaphors are conventionally noted with caps lock. In this paper we adhere to these conventions.

The Workshop on Cognitive Aspects of the Lexicon, pages 151–161
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

The fundamental bricks of WordNet architecture are represented by synsets, which can be defined as unordered sets of near-synonymic lemmas accompanied by a gloss and identified by an ID-number. Currently, the Princeton WordNet has reached version 3.1. Over time, new WordNet releases have included sets of synsets with varying IDs (for more information on the stability of these IDs over time, refer to Kafe, 2017). Synsets group together WordNet nodes, representing open-class parts of speech (lemmas) of a given language, specifically, nouns, verbs, adjectives, and adverbs. WordNet design makes use of a shallow notion of synonymy (Miller et al., 1990: 241): synsets collect synonymous word *readings* or *senses* and not "absolute synonyms", that is, words that can replace one another in all conceivable contexts (Murphy, 2010).

For example, in the current version of the Princeton WordNet, the synset "n#05022301 | the absence of heat" includes the nouns "cold", "coldness", "low temperature", "frigidity", and "frigidness". Lemmas can belong to multiple synsets, which is how WordNets represent polysemy: for example, in the Princeton WordNet, "cold" as a noun is also included in the following synsets:

- n#05733621 | the sensation produced by low temperatures
- n#14168983 | a mild viral infection involving the nose and respiratory passages (but not the lungs)

WordNets' nodes, or lemmas, are linked via lexical-morphological relations, while specific senses of lemmas, grouped in synsets, are connected through conceptual-semantic relations (for the complete set of relations in the Princeton WordNet, see Fellbaum, 1998).

Synsets, in turn, are grouped into semantic fields called 'semfields': for example, the above synsets "n#14168983 | a mild viral infection involving the nose and respiratory passages (but not the lungs)" and "n#05733621 | the sensation produced by low temperatures" belong with the semfield {Medicine and Health}, whereas the synset "n#05022301 | the absence of heat" pertains to the semfield {Physics}.

What we have discussed so far shows that, in WordNets, lexical meaning is understood as arising from relations among word senses and is accordingly stored in a relational manner. This is precisely why Geeraerts (2010) includes the WordNet project and its strands among neo-structuralist approaches to semantics, particularly among elaborations of structuralist relational semantics.

Structuralist relational semantics, best represented by Lyons' (1963), Cruse's (1986), and Murphy's (2003) seminal works, aimed to identify a theoretical apparatus and vocabulary to describe the structural relations among related words, such as synonymy, antonymy, hyponymy, hyperonymy, and meronymy, independently from encyclopedic knowledge and excluding cause-effect relations (e.g., the relation holding between *music* and *composer*). In fact, WordNets, too, fail to account for relations between concepts that are particularly close from a thematic, functional or encyclopedic point of view – a shortcoming often referred to as the 'tennis problem' (Fellbaum, 1998; Sampson, 2000). To use the tennis example, WordNets typically do not contain any coded information regarding the fact that "tennis", "ball", "racquet", and "net" are related concepts. From a taxonomic and somewhat inverse perspective, this issue is known as "IS-A overload" (Guarino, 1998; Guarino and Welty, 2002; Huang et al., 2008), a situation where semantically heterogeneous words are grouped as co-hyponyms (X is a Y) under the same hypernym. For example, the word "mask" in Princeton WordNet belongs to the synset "n#03730526 | a protective covering worn over the face" and has the following hyponyms "face mask" (for sports), "gas mask", "respirator", "gas helmet", and "welder's mask". These co-hyponyms may share the very general functionality of covering and protecting the face but are used in very different situations and belong to very different domains of reality. In cognitive semantic terms, WordNets do not capture frames (see Section 2.2).

2.2 Enhanced WordNets for Ancient Indo-European Languages

Cognitive Semantics emerged in the 1980s stemming from Cognitive Linguistics (e.g., Lakoff and Johnson, 1980; Lakoff, 1987; Taylor, 2003; Aitchinson, 2003; classic introductions in Cognitive Linguistics include Croft and Cruse, 2004; Ungerer and Schmidt, 2006; see also the scholarship overview in Geeraerts 2010: 267-272). Cognitive Linguistics looks at language in the larger context of cognition and regards language use as the essential methodological basis of linguistics. More specifically, on meaning, the three leading ideas of Cognitive Semantics can be summarized as follows:

i. Meaning exceeds the boundaries of the word and is part of larger conceptual structures, called 'frames' (Fillmore, 1975; 1985) or 'idealized cognitive models' (Lakoff, 1987), which are evoked by specific words or expressions.
ii. Meaning is contextual and pragmatically flexible, which led Cognitive Semantics to developing the idea that polysemy is structured and can be organized around a prototypical meaning (e.g., Lakoff, 1987; Brugmann, 1988; see the overview in Mangasser-Wahl, 2000) and to becoming interested in studying how actual language use drives semantic change.
iii. Expressing meaning entails perspectivization, in that complex sets of concepts, or domains, can be referred to using simpler ones, via cognitive metaphor and metonymy (Lakoff and Johnson, 1980; Kövecses, 2002; see also the handbook on metaphor theory by Gibbs, 2008).

The advantages of incorporating Cognitive Semantics into traditional lexicographic practice has been highlighted, for example, by Ostermann (2015: 48-49; earlier also by Geeraerts, 2001: 19 and by Langacker, 2005: 342), and are related to building dictionaries whose structure more closely resembles that of the mental lexicon, while simultaneously addressing the so-called 'linearization problem'. In Geeraert's (2001: 18) words, this problem describes "the fact that lexicographers [...] have to project a multidimensional clustered semantic structure onto the linear order of a dictionary". As discussed in Section 2.1, traditional WordNets clearly overcome the linearization problem, as they consist of networks of nodes linked by paradigmatic relations. However, WordNets do not store frame relations, a shortcoming that has been acknowledged as early as in Fellbaum (1998) and to which computational lexicographers still strive to find a solution (cf. Fellbaum, 2010; Koeva, 2020).

In fact, the cognitive linguistic notion of frame was extensively applied to corpus-based lexical analysis within the FrameNet project (Fillmore and Atkins, 1992; Atkins et al., 2003b; Fillmore et al., 2003, Fillmore and Petruck, 2003), whose aim is building a human- and machine- readable lexical database for English accounting for how words are used in context and how words fit into larger conceptual structures (see also https://framenet.icsi.berkeley.edu/about). Later, especially in the first decade of the 2000s, computational linguists and lexicographers attempted various computational approaches to automatically integrate WordNet and FrameNet (see, among others, Shi and Mihalcea, 2005; Tonelli and Pighin, 2009; Laparra & Rigau 2010).

Note that both WordNets' neostructuralist approach and Cognitive Semantics can be regarded as onomasiological in nature: they are both interested in looking at sets of lexical items simultaneously rather than at single lexical items. WordNet does so by clustering word senses in semantic fields ('semfields'; cf. Section 2.1), Cognitive Semantics by grouping them in frames. However, Fillmore and Atkins (1992: 76-77) well highlight the fundamental difference between these two approaches: (neo)structuralists link words, or better, word senses, directly to one another, whereas for Cognitive Semantics such relations are mediated by frames, which are made up based on our structured experience, beliefs, and practices.

In our family of WordNets of ancient languages, semantic frames are not currently annotated: thus, we have not yet incorporated in our computational lexica the first leading idea of Cognitive Semantics introduced above. This is due to the unavailability of FrameNet-like lexical databases for Latin, Ancient Greek, and Sanskrit. However, we did attempt to enhance, by adding *syntactic* frames, the verbal lemmas contained in the Ancient Greek and Sanskrit WordNets (Zanchi et al. 2021; Biagetti et al. 2023a; Biagetti et al. 2023b) and syntactic frames are being systematically annotated in the Latin WordNet in the framework of the LiLa project since earlier times (Mambrini et al. 2021).

On the other hand, in our family of WordNets we did integrate the latter two leading ideas of Cognitive Semantics presented above. First of all, to account for the contextual and pragmatical flexibility of meaning (see Section 2.2., point ii), we tag each lemma sense, that is, each synset to which each lemma belongs, for periodization, literary genre(s), and loci of attestation. For example, the Ancient Greek adjective *thermós* 'hot' is attributed to 12 synsets in the Ancient Greek WordNet. For each synset, the above pieces of information are specified as shown in the examples in (1):

(1) Lemma: *thermós*
  a. Synset: a#02407344 | having or producing a comfortable and agreeable degree of heat or imparting or maintaining heat: "a warm body"
    i. Periodization: Archaic (8th c. - 6th c. BCE); Classical (5th c. - 323 BCE); Hellenistic (323-31 BCE); Roman (31 BCE-290 CE)
    ii. Genre: Poetry, epic; Theater, comedy; Theater, tragedy; Philosophy, dialogue
    iii. Loci: Hom.*Il*.22.244; E.*Rh*.790
  b. Synset: a#01127729 | resulting from inflammation
    i. Periodization: Classical (5th c. - 323 BCE)
    ii. Genre: Theater, comedy; Theater, tragedy
    iii. Loci: S.*Ph*.696

This information makes our WordNets suitable for studies in diachronic lexicology and onomasiological variation, also in a comparative fashion (see also below about etymology). In other words, our WordNets make it possible to answer research questions such as how word meanings change over time and vary across literary genres and authors.

Moreover, each synset is tagged as literal, metonymic, or metaphoric (Figure 1), and the synset representing the prototypical meaning is identified.

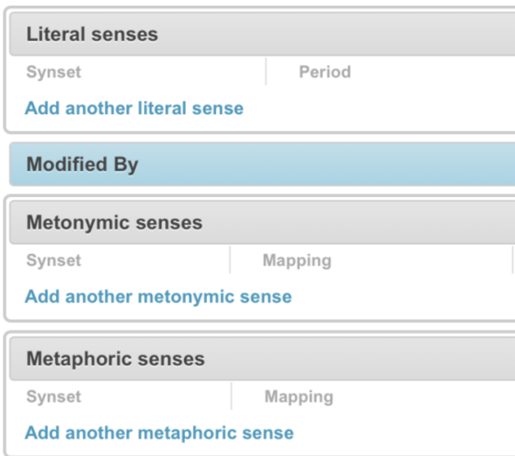


Figure 1: Fields for literal, metonymic and metaphoric senses in the annotation interface.

This type of annotation incorporates the notions of prototype and structured polysemy in our WordNets. Examples are provided in (2):

(2) Lemma: *thermós*
   a. Prototypical synset: a#02407344 | having or producing a comfortable and agreeable degree of heat or imparting or maintaining heat: “a warm body” (cf. (1)a)
   b. Literal synset: a#01195771 | used of physical heat
   c. Metaphorical synset: a#01015627 | freshly made or left: “a warm trail”

Our WordNets also contain etymological information regarding each lemma. For example, for the lemma *thermós*, the recorded Proto-Indo-European root is **g*^{hw}*er-*, to which two senses are associated, that is, the synsets in (2)a “a#02407344 | having or producing a comfortable and agreeable degree of heat or imparting or maintaining heat” and in (2)b “a#01195771 | used of physical heat”. In our WordNets, etymological information consists of the etymology proper, and optionally of an etymon, i.e., a discrete form in the history of a word’s etymological development, and one or more morphemes, i.e., discrete elements within the etymon (cf. Figure 2).

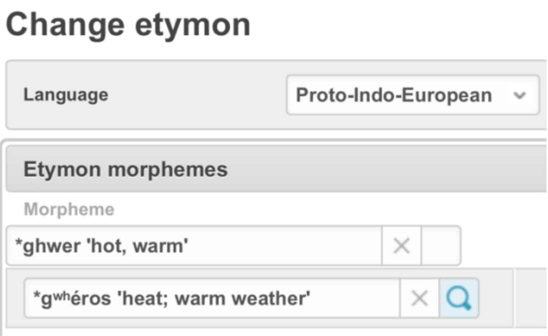


Figure 2: Morpheme and etymon annotation of PIE *g*^{wh}*éros* ‘heat; warm weather’.

In addition, our databases allow for annotating cognitive metaphors as mappings between synsets. For example, the metaphorical synset in (2)c shows that the adjective *thermós* can trigger the metaphor RECENT IS WARM. In the database, one may keep track of this metaphorical usage by linking the prototypical synset of *thermós* “a#02407344 | having or producing a comfortable and agreeable degree of heat or imparting or maintaining heat” and the metaphorical one “a#01015627 | freshly made or left”. Importantly, this relation of mapping is stored in a section of the database different from that in which traditional WordNet semantic-conceptual relations are stored. In the long run, the final goal would be to build a structured repository of conceptual metaphors (cf. MetaNet available at https://metaphor.icsi.berkeley.edu/pub/en/index.php/MetaNet_Metaphor_Wiki) for ancient Indo-European languages, as it has been partially done for Latin (Fedriani et al., 2020).
Now, before moving on to discuss the reasons why such rich annotation turned out to be problematic, which is the topic of Section 3, it is worth noting that the notion of prototype was not entirely foreign to the Princeton WordNet itself, though it was applied to a totally different part of the database. As documented in Fellbaum (1998), the Princeton WordNet deals with antonymy at a lexical level, not at a synset level. This means that, for example, it is the words “hot” and “cold” that are related by the antonymy relation rather than their corresponding synsets. In the Princeton WordNet, this way of understanding antonymy is implemented by grouping words with similar meanings in clusters (e.g., “cold,” “algid,” “chilly,” “shivery,” “frosty,” etc.), by organizing these clusters around a prototype (“cold”), and by directly linking two antonymous prototypes, while all the other members of the cluster are tagged as indirect antonyms (Figure 3). As shown in Biagetti et al. 2021, this procedure was impossible to follow for ancient languages. Thus, in our WordNets antonymy is treated at the synset level instead, apart from cases in which antonyms are morphologically derived from one another (cf. Sanskrit *uṣṇa-* ‘hot’ and *an-uṣṇa-* ‘not hot, cold’).

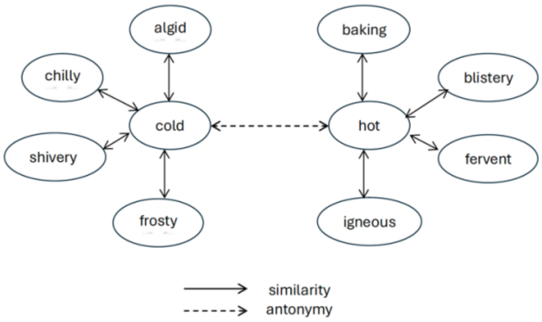


Figure 3: Bipolar Adjective Structure (adapted from Miller et al., 1993: 29).

3. A Case Study: the TEMPERATURE Domain

3.1 The Linguistics of Temperature

The linguistics of temperature is the study of how temperature concepts – e.g., HOT, COLD, LUKEWARM – are conceptualized, that is, organized in speakers’ mind and expressed in world’s languages (Koptjevskaja-Tamm, 2015: 1-40). It also explores the system of temperature terms, considering them as access points to the understanding of temperature concepts. Temperature concepts are interesting for Cognitive Semantics primarily because temperature is an invisible measure experienced by humans through their bodies and expressed through language. Consequently, temperature establishes a connection between natural phenomena, human bodies, and cognitive processes. Additionally, the perception of temperature can vary significantly, with both heat and coldness capable of being either positive or negative experiences for individuals.
The domain of TEMPERATURE is frequently employed to conceptualize more complex cognitive domains, such as the one of emotions via conceptual metaphors. For instance, AFFECTION IS WARMTH (Lakoff and Johnson, 1999: 50) and ANGER IS HEAT (Goossens, 1998; Kövecses, 2002) are two very common metaphors mapping the domain of EMOTIONS onto the one of TEMPERATURE (see also the metaphor RECENT IS WARM in example

(2)c). Being subject to the cognitive processes of metaphor and metonymy, temperature terms are therefore often highly polysemous. Thus, they constitute a good case study to present the potential and drawbacks of an annotation scheme designed to incorporate notions from Cognitive Semantics, such as the notion of structured polysemy and the distinction between literal and non-literal senses, into WordNet architecture.

3.2 Prototypical vs. secondary senses

One of the basic ideas of Cognitive Semantics is that lexical categories and polysemy networks can be thought of in terms of being structured with respect to prototypical meanings (Lakoff, 1987; Langacker, 1987: 376; see also Section 2.2). In this view, the distinct meanings or senses associated with a particular word are related in a principled way to a prototypical or sanctioning sense. According to Tyler and Evans (2003: 45-50; see also Evans, 2004: 96-98), prototypical senses are detected based on the following criteria:

i. early attestation,
ii. concreteness,
iii. predominance in the semantic network.

Importantly for our purposes, these criteria allow for detecting prototypical senses without relying on native speaker intuition.
As already mentioned in Section 2.2, in our annotation scheme, we have initially included the possibility to tag a word sense as prototypical and to distinguish it from its other literal and non-literal senses. However, the distinction between prototypical and secondary senses revealed problematic in many cases. A first problem arises if, for "early attestation," that is, Tyler and Evans' (2003) criterion i), we understand the sense reconstructed for the Proto-Indo-European root associated with the lemma-node in question. In fact, as we have seen for **g^{hw}er-* above, the etymology of a lemma can have more than one sense, which would make the choice of one prototypical sense arbitrary. Even if we interpret the criterion of early attestation as referring to the oldest attested sense in the language under scrutiny, the distinction is problematic. Take, for example, the Ancient Greek noun *págos*. For this noun, the Liddell-Scott-Jones dictionary (i.e., the reference dictionary of Ancient Greek) provides the following definition:

(3) *that which is fixed or firmly set*:
i. crag, rock, generally rocky hill (often used in conjunction with Arēs to mean the Areopagus at Athens)
ii. after Homer = *pagetós*, frost

Chantraine's *Dictionnaire étymologique de la langue grecque* includes *págos* among the derivatives of the verb *pḗgnumi* 'make fast' and attributes to it the meaning 'that which is fixed, hard'. From this derives the meaning 'rock, cliff', attested in older sources and retained in Attic in the name of the Areopagus (*Áreios Págos*) and, after Homer, the meaning 'frost, cold'. Other derivatives of *pḗgnumi* listed by Chantraine are *pagetós* 'frost', *pagerós* 'frosty, cold', *págiōs* 'firm, solid'; from the stem *pēg-*, we find *pēgós* 'solid, vigorous', *pēgás* 'hoar-frost, rime', and others, all suggesting a connection between 'firmness' and 'coldness'.
The *Brill Etymological Dictionary of Ancient Greek* (Beekes and van Beck, 2010) states that *págos* is derived from *págē* [f.] 'snare, trap, anything that fixes'. This dictionary does not provide any specific meaning for *págos* but, like the Liddell-Scott-Jones, asserts that its meaning is equivalent to that of *pagetós* 'frost'. The case of *págos* makes it clear that the criteria for identifying a prototypical sense can conflict with each other and prevent annotators from selecting one. Indeed, the earlier attestation of the sense in (3)i, 'crag, rock', instantiated by example (4), would lead to its selection as the prototypical sense (synset "n#06669293 | a lump of hard consolidated mineral matter"). However, this sense is employed in very specific contexts, often indicating the Areopagus in Athens (5). On the other hand, the higher frequency of sense in (3)ii, 'frost' shown in (6), suggests that this might be selected as the prototypical sense (synset "n#09741425 | the formation of frost or ice on a surface").

(4) *ou gàr ésan liménes*
NEG PTC be:IMPF.3PL harbor:NOM.PL
nēôn ókhoi,
ship:GEN.PL shelter:NOM.PL
oud' epiōgaí
NEG roadstead(F):NOM.PL
all' aktaì problêtes
but headland(F):NOM.PL projecting(F):NOM.PL
ésan spiládes te págoi te
be:IMPF.3PL reef(F):NOM.PL and rock:NOM.PL and
'for there were neither harbors where ships might ride, nor road-steads, but projecting headlands, and reefs, and cliffs' (Hom.*Od*.5.404-405)

(5) *hoi dè Pérsai*
DET PTC persian:NOM.PL
hizómenoi epì tòn
place:PTCP.PL upon DET
katantíon tês akropólios
over.against DET acropolis(F):GEN
ókhthon, tòn athēnaîoi
hill:ACC DET athenian:NOM.PL
kaléousi arḗion ***págon***
call:3PL of.ares:ACC **rock:ACC**
'The Persians took up a position on the hill opposite the acropolis, which the Athenians call the Areopagus.' (Hdt.8.52.3)

(6) *pou* ***págou*** *khuthéntos,*
where **frost:GEN** spread:AOR.PTCP
oîa kheímati,
such.as winter(N):DAT
xúlon ti thraûsai,
firewood(N):NOM any break:AOR.INF
taût' àn exérpōn
this PTC creep.out:PTCP
tálas emēkhanōmēn
wretched:NOM manage:IMPF.1SG

'if when the frost had spread, as often happens in winter, a bit of firewood had to be broken, I would creep out in pain and manage it.' (S.*Ph*.293)

Given the opaque nature of the difference between prototypical and non-prototypical senses, we decided to forgo accounting for such a distinction in our WordNets.

3.3 Literal vs. Non-literal Senses

The third basic idea of Cognitive Semantics is that complex sets of concepts can be referred to using simpler ones, through conceptual metaphor and metonymy. For this reason, in our WordNets, we have decided to distinguish between the literal, metaphorical, and metonymic senses of a word (see Section 2.2).

However, even a seemingly straightforward distinction like that between literal and non-literal senses proves problematic in some cases. One of these cases is when the sense(s) of a lemma can be analyzed diachronically as derived from a simpler or more concrete sense through metaphorical or metonymic processes, but such simpler sense is not attested in the history of the language. For example, the Latin verb *ferveo* (or *fervo*) 'seethe, boil' and *furo* 'rage, be furious' go back to the same Proto-Indo-European root **b^heru̯-* 'seethe, boil', via two allomorphs, *feru(e)-* and *fur(o)-*, which were reassigned to separate paradigms (Kölligan, 2020). The link between boiling and rage is licensed by the conceptual metaphor ANGER IS A HOT FLUID IN A CONTAINER (Kövecses, 2010: 123; cf. also Lakoff, 1987: 383), which in turn derives from the more general ANGER IS HEAT via the BODY IS A CONTAINER metaphor. In Latin, *ferveo* retained both the literal (7)a and the metaphorical (7)b senses, whereas *furo* is primarily employed in the metaphoric sense of 'rage, be furious' shown in (8)a. However, *furo* is also found in contexts such as (8)b, where *furit* may in fact refer to magma of the volcano boiling underneath the earth. Finally, the deverbal noun *furor* 'wrath' only features the metaphorical meaning (9).

(7) *ferveo* 'seethe, boil' but also 'be angry, rage':

a. literal meaning

quin *omnia* *malit*
COMPL all:ACC.PL prefer:SBJV.3SG
quaecumque *inmundis* ***fervent***
REL.NOM.PL nasty:ABL.PL be.hot:3PL
allata *popinis.*
bring:PTCP.PRF.NOM.PL eating.house:ABL.PL

'It (the stomach) will prefer everything which is brought smoking hot from the nasty eating-houses.' (Hor.*Sat*.2.4.61-62)

b. metaphorical meaning

animus *tumida*
heart:NOM swelling:ABL
feruebat *ab* *ira*
be.hot:IMPF.3SG from anger:ABL

'His heart became hot with swelling anger.' (Ov.*Met.*2.602)

(8) *furo* 'rage, be furious'

a. metaphorical meaning

quo *genere* *Athamantem*
REL.ABL sense:ABL Athamas:ACC
Alcmaeonem *Aiacem*
Alcmaeon:ACC Ajax:ACC
Orestem ***furere*** *dicimus …*
Orestes:ACC be.furious:INF say:PRS.1PL

'[the mind is influenced … by the stronger power of wrath or fear or pain,] in the sense in which we say that Athamas, Alcmaeon, Ajax and Orestes **are furious**.' (Cic.*Tusc.*3.11)

b. literal meaning

ex *imis* *uero*
from more.profound:ABL.PL indeed
furit *ignibus* *impetus*
rave:3SG fire:ABL.PL attack:NOM
Aetnae
Aetna:GEN

'the impetuous Aetna **raves** indeed from more profounder fires.' (Lvcr.593)

(9) *furor* 'wrath'

a. metaphorical meaning

cum ***caeci*** ***furore*** *in*
when blind:NOM.PL rage:ABL into
uolnera *ac* *ferrum*
wound:ACC.PL and sword:ACC
uecordi *audacia*
reckless:ABL daring:NOM
ruerent
rush:SBJV.IMPF.3PL

'when they (the Astapans), **blind with rage**, rushed upon wounds and the sword with reckless daring.' (Liv.28.22.14)

Given the situation presented above, *ferveo* should be annotated in our WordNet as having a literal sense 'seeth, boil' (synset "v#00261276 | bring to, or maintain at, the boiling point") and a metaphorical one 'rage' ("v#01225618 | feel intense anger"), as both are attested in the history of Latin. In the case of *furor*, the sense 'wrath' ("n#05588321 | intense anger") should be annotated as the literal one, as it is the only one attested in the texts. Finally, although we know that 'rage, be furious' is the result of a metaphorical shift, we should tag this sense as the literal sense of *furo*, as this is the primary meaning attested in the texts; cases like (8)b, on the other hand, can be seen as instances of personification, corresponding to English *angry sea* and belonging to the same personification process as Latin *mare placidum* 'calm sea'.

The case of Latin *furo* and *furor* is different, for example, from that of Sanskrit *ghr̥ṇā-* (Proto-Indo-European **g^{hw}er-* 'burn'; the Sanskrit root *ghar*, from which *ghr̥ṇā-* is derived, is not attested with verbal use; cf. EWA s.v.): for this noun, too, we know that from the literal sense 'heat' a metaphorical sense 'compassion' has been derived (through the metaphor AFFECTION IS WARMTH), which has then completely replaced the former sense. However, in this case, both meanings are attested in the history of the language, the former in Vedic Sanskrit shown in (10) and the latter in Epic and Classical Sanskrit shown in (11). Therefore, in the Sanskrit WordNet, we annotate ‘heat’ (“n#07805780 | a form of energy that is transferred by a difference in temperature”) as a literal sense and ‘compassion’ (“n#05615476 | a deep awareness of and sympathy for another's suffering”) as a non-literal sense of *ghr̥ṇā-*.

(10) *párīm̐* ***ghr̥ṇā́*** *carati*
around heat(F):NOM go:3SG
titviṣé *śávo*
flare:PF.3SG.MID power(N):NOM
‘Glowing heat encircles him [=Indra], and his vast power flared.’ (*Ṛgveda* 1.52.6a)

(11) *ahiṃsā* *satya-vacanam*
non-violence(F):NOM sincere-word(N):NOM
ānṛśaṃsyam *dama*
kindness(N):NOM control:NOM
ghṛṇā
compassion(F):NOM
‘non-violence, sincere word, kindness, control, compassion’ (*Mahābhārata* 12, 80, 17.1; from de Rossi 2023: 93)

3.4 Metaphoric vs. Metonymic Senses

There are many cases where distinguishing between metaphor and metonymy becomes challenging, especially considering the standard definitions of metaphor and metonymy (cf. Goossens, 1990). As we have seen in Section 2.2, metaphor consists in conceptualizing one domain in terms of another (Lakoff and Johnson, 1980); in metonymy, an element within a domain provides access to another element within the same domain (Kövecses and Radden, 1998; Radden and Kövecses, 1999). Issues in distinguishing metaphor from metonymy arise in cases where it is unclear whether we are dealing with one domain or two, and this happens because some metaphors derive from metonymies (Kövecses, 2013: 78).

Take for instance the metaphor ANGER IS HEAT. In our folk model of emotions, the latter are seen as resulting in some physiological effects. Since anger is often accompanied by an increase in body temperature, sweating, and facial flushing, the conceptualization of anger in terms of increased body heat is licensed by the metonymy EFFECT FOR CAUSE. Paraphrasing it as THE PHYSIOLOGICAL EFFECTS OF AN EMOTION ARE THE EMOTION ITSELF, it becomes clear that the sole domain of EMOTIONS is involved here, and so we are dealing with a metonymy.

The conceptual metaphor ANGER IS HEAT arises from the cognitive process of generalization (Kövecses and Radden, 1998: 61; Kövecses, 2013: 80). If body heat is generalized into heat, a second domain, the one of TEMPERATURE, comes into play and becomes the source domain of a metaphor. In example (12), the verb *dah-* ‘burn’ is employed with reference to ‘anger’ (synset “v# 01248170 | feel strong emotion, esp. anger or passion”). Since this verb is usually referred to the burning of fire, and not to personal feeling temperature, we should probably annotate its use in (12) as metaphorical.

(12) *amarṣeṇa* *susampūrṇaḥ*
anger:INS fill:PTCP.PASS.PST.NOM
dahyamānaḥ *divāniśam*
burn:PTCP.PASS.NOM night.and.day
‘I am full of anger, I burn day and night.’ (*Mahābhārata* 2,43.21.1; from de Rossi 2023: 65)

Another emotion that is often conceptualized in terms of warmth is love, or romantic passion (13). Since love does not cause an increase in body temperature – though blushing may be an effect of it – the association appears to be an instance of a more general metaphor A STRONG EMOTION IS HEAT. From this it follows that LACK OF HEAT IS LACK OF EMOTION, and consequently LACK OF HEAT IS LACK OF LOVE (14).

(13) *et* *amore* ***ardeo***
and love:ABL burn:1SG
‘And I burn with passion.’ (Ter.*Eun*.72)

(14) ***tepida*** *mens*
warm:NOM mind:NOM
‘Cooled heart.’ (Ov.*Ars*.2, 445)

The examples above demonstrate the difficulties that are often encountered in distinguishing between metaphor and metonymy, in general. This is especially true for ancient languages like those represented in our WordNets, for which we cannot rely on native speaker intuition to reconstruct the cognitive processes that license the non-literal meanings of a word.

4. Conclusions

In this paper, we have presented some challenges encountered in combining neo-structuralist and cognitive approaches to semantics for building WordNets of ancient languages. Indeed, although the architecture of our Ancient Greek, Latin, and Sanskrit WordNets follows the one of the original Princeton WordNet in many respects, we integrate notions of Cognitive Semantics.

This approach seems promising as it allows for a more fine-grained and “more realistic” representation of meaning, and thus facilitates studies in diachronic semantics and lexicology.

However, an initial annotation phase in our project has shown that some of the integrations initially planned can hardly be implemented in the relational database behind our WordNets. These difficulties partly arise from dealing with ancient languages for which we lack native speakers to judge the validity of our analyses. Other challenges stem from the need to achieve maximum granularity in meaning representation while maintaining a sustainable workflow for annotators, who must be provided with guidelines that are clear and valid for the majority of lemmas they annotate.

For these reasons, we decided to forgo accounting for the distinction between the prototypical sense, that is, the sense that speakers identify as the most representative of a lexical unit (Evans, 2004: 92), and other secondary senses. In addition to the previously mentioned lack of native speakers for these languages, the difficulty in drawing such a distinction arises from the fact that the criteria for identifying a prototypical sense can conflict with one another, as demonstrated in Section 3.2 for Ancient Greek *págos* 'rock', 'frost'.

For the distinction between literal and non-literal senses, it is crucial to keep separate the senses reconstructed by the etymology of a lexeme from the ones actually attested in the history of the language. As a result, we treat Latin *furor* differently from Sanskrit *ghṛṇā-*. For *furor*, we can hypothesize that from a literal meaning related to 'boiling' (from the Proto-Indo-European root **bʰeru̯-*), a metaphorical meaning 'anger' has developed. However, since the latter is the only sense attested in the history of Latin, we must annotate it as literal. Similarly, for *ghṛṇā-* we know that from a literal meaning 'heat', a metaphorical meaning 'compassion' has developed, which then replaced the former. However, since both senses are attested in the history of the language, we annotate the former as literal and the latter as non-literal.

Finally, given the close relationship between metonymy and metaphor in some cases, it is not always possible to distinguish senses derived through one or the other cognitive process. Moreover, even if an in-depth study of a given semantic field allowed for an agreement on what is metonymic and what is metaphorical, such a workflow would not be sustainable for annotators, who are primarily BA and MA students in Classics and Linguistics, and not even for their supervisors, who would need to double-check the most complex cases one by one. On the one hand, the sometimes-blurry distinction would result in a low inter-annotator agreement; on the other hand, since dictionaries do not contain all the necessary information to maintain this distinction, annotators would have to look at individual examples in context, which is a very time-consuming process. For these reasons, in the next phases of the project, we limit ourselves to the distinction between literal and non-literal senses.

5. Bibliographical References

Aitchison, J. (2003). *Words in the Mind: An Introduction to the Mental Lexicon*. Blackwell, Oxford, 3rd edition.

Atkins, B.T.S., Fillmore, C. and Johnson, C. (2003a). *Lexicographic Relevance: Selecting Information from Corpus Evidence*. International Journal of Lexicography, 16/3, pp. 251–280.

Atkins, B.T.S, Rundell, M. and Sato, H. (2003b). *The Contribution of FrameNet to Practical Lexicography*. International Journal of Lexicography, 16/3, pp. 333–357.

Beekes R. S. P. and van Beck, L. (2010). *Etymological Dictionary of Greek*. Brill, Leiden.

Biagetti, E., Zanchi, C., and Short, W. M. (2021). Toward the creation of WordNets for ancient Indo-European languages. In P. Vossen and C. Fellbaum (Eds.), *Proceeding of the 11th Global WordNet Conference*, University of South Africa (UNISA): Global Wordnet Association, pp. 258-266.

Biagetti E., Zanchi, C., and Luraghi, S. (2023a). Linking the Sanskrit WordNet to the Vedic Dependency Treebank: a pilot study. In *Proceedings of the 12th Global Wordnet Conference*, pages 77–83, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Biagetti, E., Brigada Villa, L., Zanchi, C., and Luraghi S. (2023b). Enhancing the semantic and conceptual description of Ancient Greek verbs in WordNet with VerbNet and FrameNet: a treebank-based study. In *Papers from the Annual International Conference "Dialogue"*, Vol. 22 (Supplementary volume), pages 1009–1020.

Bizzoni Y., Boschetti, F., Diakoff, H., Del Gratta, R., Monachini, M., and Crane, G. (2014). The Making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1140–1147, Reykjavik, Iceland. European Language Resources Association (ELRA).

Boschetti, F. (2019). Semantic Analysis and Thematic Annotation. In M. Berti (Ed.), *Digital Classical Philology*, Berlin: De Gruyter, pp. 321-339.

Brugman, C. (1988). *The Story of Over: Polysemy, Semantics and the Structure of the Lexicon*. Garland, New York.

De Rossi, N. (2023). *"Bruciare" nel Mahābhārata: un'analisi delle radici dah-, tap-, uṣ-, jval- (e jvar-), ghṛ- e śuc- nel quadro della semantica cognitive*. MA Thesis, University of Pavia.

Croft, W. and Cruse, D. A. (2004). *Cognitive Linguistics*, Cambridge University Press, Cambridge.

Cruse, D. A. (1986). *Lexical Semantics*, Cambridge University Press, Cambridge.

Evans, V. (2004). The structure of time. The Structure of Time, 1-292.

Fedriani C., De Felice, I., and Short, W. M. (2020). The digital Lexicon Translaticium Latinum: Theoretical and methodological issues. In C. Marras, M. Passarotti, G. Franzini and E. Litta (Eds.), *Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). La svolta inevitabile: Sfide e prospettive per l'informatica umanistica*, pp.106-112. https://lexelat.unige.it/sites/lexelat.unige.it/files/pagine/PDF.pdf

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA, MIT Press.

Fellbaum, C. (2010). Harmonizing WordNet and FrameNet. In H. Loftsson, E. Rögnvaldsson, and S. Helgadóttir (Eds.), *Advances in Natural Language Processing. NLP 2010. Lecture Notes in Computer*

Science(), vol 6233. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14770-8_2.

Fillmore, C. J. (1975). An alternative to checklist theories of meaning. In C. Cogen, H. Thompson, G. Thurgood, K. Whistler, and J. Wright (Eds.), *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*, Berkeley, CA: Berkeley Linguistics Society, pp. 123–131.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica* 6: 222–254.

Fillmore, C. J., and Atkins, B. T. (1992). Toward a Frame-Based Lexicon: The Semantics of RISK and Its Neighbors. In A. Lehrer and E. F. Kittay (Eds.), *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 75–102.

Fillmore, C., Johnson, C. and Petruck, M. (2003). *Background to FrameNet*. International Journal of Lexicography, 16/3, pp. 235–250.

Fillmore, C. and Petruck, M. (2003). *FrameNet Glossary*. International Journal of Lexicography, 16/3, pp. 359–261.

Franzini, G., Peverelli, A., Ruffolo, P., Passarotti, M. C., Sanna, H., Signoroni, E., Ventura, V., and Zampedri, F. (2019). Nunc Est Aestimandum: Towards an Evaluation of the Latin WordNet. In *Proceedings of the Sixth Italian Conference on Computational Linguistics,* (BARI -- ITA, 13-15 November 2019), Accademia University Press, Torino, -- ITA 2019:<<COLLANA DELL'ASSOCIAZIONE ITALIANA DI LINGUISTICA COMPUTAZIONALE>>, 1-8.

Geeraerts, D. (2001). The definitional practice of dictionaries and the Cognitive Semantic conception of polysemy. *Lexicographica* 17: 6–21.

Geeraerts, D. (2007). Lexicography. In D. Geeraerts and H. Cuyckens, (Eds.), *The Oxford Handbook of Cognitive Linguistics*, New York: Oxford University Press, pp. 1160-1175.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford University Press, Oxford.

Gibbs, Jr. R. W. (2008). *The Cambridge Handbook of Metaphor and Thought*. Cambridge University Press, Cambridge.

Goossens, L. (1990). Metaphtonymy: the interaction of metaphor and metonymy in expressions for linguistic action. *Cognitive Linguistics* 1-3: 323–340.

Goossens L. (1998). Meaning extension and text type. *English Studies* 79:2: 120-143. DOI: 10.1080/00138389808599120.

Guarino, N. (1998). Some Ontological Principles for Designing Upper Level Lexical Resources. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada (Eds.), *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 527-534.

Guarino, N. and Welty, C. (2002). Identity and Subsumption. In R. Green, C. A. Bean, and S. H. Myaeng (Eds.), *The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management*, Verlag: Springer, pp. 111-125.

Hellwig, O. (2017). Coarse Semantic Classification of Rare Nouns Using Cross-Lingual Data and Recurrent Neural Networks. In C. Gardent and C. Retoré (Eds.), *Proceedings of the 12th International Conference on Computational Semantics (IWCS)* https://aclanthology.org/W17-6811/

Huang, C.-R., Su, I., Hsiao, P.-Y., Ke, X.-L. (2008). Paranymy: Enriching Ontological Knowledge in WordNets. *Proceedings of the Fourth Global WordNet Conference*, pages 221–228, Szeged, Hungary, Juhász Press Ltd.

Kafe, E. (2017). How Stable are WordNet Synsets?. *LDK Workshops* 2017: 113-124. https://ceur-ws.org/Vol-1899/CfWNs_2017_proc1-paper_1.pdf

Khan, F., Minaya Gómez, F. J., Cruz González, R., Diakoff, H., Díaz-Vera, J. E., McCrae J. P., O'Loughlin, C., Short W. M., and Stolk, S. (2022). Towards the Construction of a WordNet for Old English. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3934–3941, Marseille, France. European Language Resources Association.

Khan, F., McCrae, J. P., Minaya Gómez, F. J., Cruz González, R., and Díaz-Vera, J. E. (2023). Some Considerations in the Construction of a Historical Language WordNet. In *Proceedings of the 12th Global Wordnet Conference*, pages 101–105, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Koeva, S. (2020). Semantic Relations and Conceptual Frames. In S. Koeva (Ed.), *Towards a Semantic Network Enriched with a Variety of Semantic Relations*, Sofia: Professor Marin Drinov Publishing House of BAS, pp. 7-20.

Kölligan, D. (2020). Seething anger: Latin furor. In L. Repanšek, H. Bichlmeier, and V. Sadovski (Eds.), *vácamsi miśrā́ kr̥ṇavāmahai, Proceedings of the internatioonal conference of the Society for Indo-European Studies and IWoBA XII*, Ljubljana 4-7 June, 2019, Hamburg: Baar, pp. 397-412.

Koptjevskaja-Tamm, M. (2015). *The linguistics of temperature*. Benjamins, Amsterdam/Philadelphia

Kövecses, Z. (2002) Metaphor: A Practical Introduction. Oxford University Press, Oxford.

Kövecses, Z. (2010). *Metaphor. A Practical Introduction*. Oxford University Press, Oxford, 2nd edition.

Kövecses, Z. (2013) The Metaphor–Metonymy Relationship: Correlation Metaphors Are Based on Metonymy. *Metaphor and Symbol* 28(2): 75-88.

Kövecses, Z and Radden, G. (1998). Metonymy: Developing a cognitive linguistic view. *Cognitive Linguistics* 9(1): 37–77.

Lakoff, G. (1987). *Women, Fire and Dangerous Things. What categories reveal about the mind*, University of Chicago Press, Chicago/London.

Lakoff, G., and, Johnson, M. (1980). *Metaphors we live by.* University of Chicago Press, Chicago.

Lakoff, G., and, Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenges to Western Thought*. University of Chicago Press, Chicago.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar 1: Theoretical Prerequisites*. Stanford University Press, Stanford, CA.

Langacker, R. W. (2005). Cognitive Grammar: the State of the Art and Related Issues (An Interview with Ronald Langacker by József Andor). *Acta Linguistica Hungarica* 52(4): 341–366.

Laparra, E. and Rigau, G. (2010). eXtended WordFrameNet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 10), Valletta, Malta. European Language Resources Association (ELRA). [http://www.lrec-conf.org/proceedings/lrec2010/pdf/799_Paper.pdf].

Lyons, J. (1963). *Structural Semantics*, Blackwell, Oxford.

Mambrini, F., Passarotti, M. C., Litta Modignani Picozzi, E. M. G., Moretti, G. (2021). Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin, in Further with Knowledge Graphs. In *Proceedings of the 17th International Conference on Semantic Systems*, 6-9 September 2021, Amsterdam, The Netherlands, (AMSTERDAM -- NLD, 06-09 September 2021), IOS Press, AMSTERDAM -- NLD 2021: 53 16-28. [10.3233/SSW210032] [http://hdl.handle.net/10807/183431].

Mangasser-Wahl, M. (2000). *Prototypentheorie in der Linguistik: Anwendungsbeispiele – Methodenreflexion – Perspektiven*. Stauffenburg. Tübingen.

Miller, G. A. and Fellbaum, C. (2007). WordNet then and now. *Language Resources and Evaluation* 41: 209–214.

Miller G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography* 3(4): 235-44.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A Semantic Concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*. https://aclanthology.org/H93-1061.pdf.

Minozzi, S. (2017). Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'Information Retrieval. In P. Mastandrea (Ed.), *Strumenti digitali e collaborativi per le Scienze dell'Antichità*, Venezia: Edizioni Ca' Foscari, pp. 123-134.

Murphy, M. L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy, and Other Paradigms*. Cambridge University Press, Cambridge.

Murphy, M. L. (2010). *Lexical meaning*. Cambridge University Press, Cambridge, UK. DOI: 10.1017/CBO9780511780684.

Ostermann, C. (2015). *Cognitive Lexicography: A New Approach to Lexicography Making Use of Cognitive Semantics*. De Gruyter, Berlin.

Radden, G. and Kövecses, Z. (1999). Towards a Theory of Metonymy. In P. Klaus-Uwe and R. Günter (Eds.), *Metonymy in Language and Thought.* Amsterdam/ Philadelphia: John Benjamins, pp. 17–59.

Sampson, G. (2000). Review of WordNet: An Electronic Lexical Database. *International Journal. of Lexicography* 13.54–9.

Shi, L. and Mihalcea, R. (2005). Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In A. Gelbukh, A. (Eds), *Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science*, vol 3406. Berlin / Heidelberg: Springer, pp. 100–111. [https://doi.org/10.1007/978-3-540-30586-6_9].

Taylor, J. R. (2003). *Linguistic Categorization*. Oxford University Press, Oxford, 3rd edition.

Tonelli, S. and Pighin, D. (2009). New Features for FrameNet – WordNet Mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (CoNLL), pages 219–227, Boulder, Colorado, June 2009. Association for Computational Linguistics.

Tyler, A. and Evans, V. (2003). *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning and Cognition*. Cambridge University Press, Cambridge.

Ungerer, F. and Schmid, H.-J. (2006). *An Introduction to Cognitive Linguistics.* Longman, London, Second edition.

Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic, Dordrecht.

Vossen, P. (2004). EuroWordNet: a multilingual database of autonomous and language specific wordnets connected via an inter-lingual index. *International Journal of Lexicography* 17: 161–73.

Zanchi C., Luraghi, S., and Biagetti, E. (2021). Linking the Ancient Greek WordNet to the Homeric Dependency Lexicon. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, Vol. 20. [https://www.dialog21.ru/media/5555/zanchicpluslu raghilplusbiagettie029.pdf].

6. Language Resource References

Zanchi C., Luraghi, S., and Biagetti, E. (2021). Linking the Ancient Greek WordNet to the Homeric Dependency Lexicon. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, Vol. 20. https://www.dialog21.ru/media/5555/zanchicpluslur aghilplusbiagettie029.pdf.

Fellbaum, C. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. https://wordnet.princeton.edu, accessed February 28, 2024.

Ruppenhofer, J., Ellsworth, M. Petruck, M. R. L., Johnson, C. R., Baker, C. F.; Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice* (revised ed.). Berkeley, CA: International Computer Science Institute. http://framenet.icsi.berkeley.edu, accessed February 28, 2024.

7. Acknowledgements

Research for this paper and for the creation of the Ancient Greek, Latin and Sanskrit WordNets has been supported by European Union funding – NextGenerationEU – Missione 4 Istruzione e ricerca - componente 2, investimento 1.1" Fondo per il Programma Nazionale della Ricerca (PNR) e Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN)" project 2022YAPFNJ “Linked WordNets for Ancient Indo-European Languages”; CUP F53D2300490 000.

This paper results from the sustained collaboration of all authors. For Italian academic purposes, Erica Biagetti is responsible of Section 3, Chiara Zanchi is responsible of Section 2, whereas all authors are responsible of Sections 1 and 4.

8. Appendix A: abbreviations used in the glosses

The interlinear glosses used in the examples follow the Leipzig glossing rules (https://www.eva.mpg.de/lingua/resources/glossing-rules.php).

1	first person
2	second person
3	third person
ABL	ablative
ACC	accusative
AOR	aorist
COMPL	completive
DAT	dative
DET	determiner
F	feminine
GEN	genitive
IMPF	imperfect
INF	infinitive
INS	instrumental
MID	middle
N	neuter
NEG	negation
NOM	nominative
PASS	passive
PL	plural
PRF	perfect
PST	past
PTC	particle
PTCP	participle
REL	relative
SG	singular
SBJV	subjunctive

In glosses, the nominal number is specified only if it is plural or dual (singular is not indicated); similarly, gender is specified only if it is feminine or neuter (masculine is not indicated). Among verbal categories, present tense, indicative mood, and active voice are likewise not indicated.

9. Appendix B: authors and works cited in the examples

The abbreviations used in this paper are taken from the *Thesaurus Linguae Graecae* (https://stephanus.tlg.uci.edu/lsj/01-authors_and_works.html) for Ancient Greek examples and from the *Thesaurus Linguae Latinae* (https://thesaurus.badw.de/en/tll-digital/index/a.html) for Latin ones.

Cic. = M. Tullius Cicero
    Tusc. = Tusculanae disputationes
E. = Euripides Tragicus
    Rh. = Rhesus
Hdt. = Herodotus Historicus, Storiae
Hom. = Homerus Epicus
    Od. = Odyssea
Hor. = Q. Horatius Flaccus
    Sat. = Saturae (sermones)
Liv. = T. Livius Patavinus, *Ab urbe condita*
Lvcr. = T. Lucretius Carus, *De rerum natura*
Ov. = P. Ovidius Naso
    Ars. = Ars amatoria
    Met. = Metamorphoses
S. = Sophocles Tragicus
    Ph. = Philoctetes
Ter. = P. Terentius Afer
    Eun. = Eunuchus

SensoryT5: Infusing Sensorimotor Norms into T5 for Enhanced Fine-grained Emotion Classification

Yuhan Xia[1], Qingqing Zhao[2], Yunfei Long[1], Ge Xu[3], Jia Wang[4]
[1]School of Computer Science and Electronic Engineering, University of Essex, UK
[2]Institute of Linguistics, Chinese Academy of Social Sciences, China
[3]College of Computer and Control Engineering, Minjiang University, China
[4]Department of Intelligent Science, Xi'an Jiaotong-Liverpool University, China
{yx23989, yl20051}@essex.ac.uk
zhaoqq@cass.org.cn, xuge@pku.edu.cn, Jia.Wang02@xjtlu.edu.cn

Abstract

In traditional research approaches, sensory perception and emotion classification have traditionally been considered separate domains. Yet, the significant influence of sensory experiences on emotional responses is undeniable. The natural language processing (NLP) community has often missed the opportunity to merge sensory knowledge with emotion classification. To address this gap, we propose SensoryT5, a neuro-cognitive approach that integrates sensory information into the T5 (Text-to-Text Transfer Transformer) model, designed specifically for fine-grained emotion classification. This methodology incorporates sensory cues into the T5's attention mechanism, enabling a harmonious balance between contextual understanding and sensory awareness. The resulting model amplifies the richness of emotional representations. In rigorous tests across various detailed emotion classification datasets, SensoryT5 showcases improved performance, surpassing both the foundational T5 model and current state-of-the-art works. Notably, SensoryT5's success signifies a pivotal change in the NLP domain, highlighting the potential influence of neuro-cognitive data in refining machine learning models' emotional sensitivity.

Keywords: emotion classification, sensory information, attention mechanism, pre-trained language model

1. Introduction

Affective computing stands at the intersection of technology and human emotions (Li et al., 2017), whereby sentiment analysis and emotion recognition are generally merged to give machines a semblance of human-like emotional understanding. Specifically, sentiment analysis (SA) seeks to decode the attitudes and viewpoints of opinion holders using computational methods (Lu et al., 2023), providing a coarse-grained categories of polarities: positive, negative, or neutral (Long et al., 2019b). Driven by recent advancements in deep learning and bolstered by vast labeled datasets, discriminating sentiments in standard contexts has become progressively more tractable. Cutting-edge models, including the likes of BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), and the T5 (Raffel et al., 2020) series, have consistently set benchmarks, achieving high accuracies on an array of sentiment classification tasks.

By contrast, emotion analysis (EA) has received less notable results in recent years. One of the reasons is that different from SA offering a coarse-grained outlook, EA paints a detailed picture. That is, EA not only distinguishes between basic sentiments but also identifies nuanced emotions such as joy, anger, sadness, surprise, and among others (Ekman, 1992). Thus, the task of EA is complicated by the sheer variety of emotional categories. For instance, distinguishing closely related emotions like "contentment" and "happiness" or "annoyance" and "anger" requires a discerning approach, especially when the medium is textual content. Thus, this study introduces a SensoryT5 model, tailored to infuse sensory data, which is cognitively more related to emotions and includes linguistically more enriched features, into neural architectures, to achieve a profound comprehension of emotions.

The relationship between emotion and perception/sensation has been verified repeatedly in various disciplines. From a neuroscientific perspective, emotion and sensory information are processed in an overlapping neural region, i.e., the amygdala (Šimić et al., 2021). Shifting the lens to psychology, emotion and perception are intertwined (Zadra and Clore, 2011). For example, the sense of taste shows an inherent link with reward and aversion mechanisms, such as sucrose being perceived as sweet and desirable, whereas quinine being recognized as bitter and repulsive (Yamamoto, 2008). In addition, emotion as a kind of interoception forms an indispensable part of human sensations, when a wide definition of sensory perception adopted (Connell et al., 2018; Lynott et al., 2020). In terms of the linguistic conceptualization of emotions, people more frequently use figurative language instead of literal emotion terms to convey emotions (Fainsilber and Ortony, 1987; Lee, 2018), and the conceptual metaphor EMOTION IS PERCEPTION is grounded

The Workshop on Cognitive Aspects of the Lexicon, pages 162–174
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

in abundant language usages to show that the human senses are fruitful sources for verbalizing emotions (e.g., sweet and bitter) (Lakoff and Johnson, 1980; Kövecses, 2019; Müller et al., 2021).

Given the intertwined relation between emotion and perception/sensation, this study posits that incorporating sensory information into a computational framework can capture the nuanced interplay between them, hence offering a reflection of intricate human affective understanding. Specifically, we utlize the Lancaster Sensorimotor Norms (Lynott et al., 2020), which include language-specific lexical properties representing the correlation between conceptualized lexical meanings and sensory modalities.

Our work boasts three pivotal advancements: (1) We introduce SensoryT5, an innovative architecture that enhances transformer-based fine-grained emotion classification models by seamlessly embedding sensory knowledge. Marking one of the pioneering endeavors, SensoryT5 is adapted at harmonizing both the nuances of contextual attention and the intricacies of sensory information-based attention. (2) The SensoryT5 leverages sensorimotor norms within transformer text classification frameworks, contributing to the ongoing efforts to incorporate neuro-cognitive data in NLP tasks. Thus, our work not only demonstrates the practical benefits of this integration in improving emotion classification tasks, but also encourages continued interdisciplinary dialogue and research between the domains of language processing and neuro-cognitive science. (3) Assessments across multiple real-world datasets pertinent to fine-grained emotion classification affirm that our approach amplifies the efficacy of pre-existing models considerably, even surpassing contemporary state-of-the-art methodologies on selected datasets. This endeavor underscores the value of cognition-anchored data in sculpting attention models. Our findings illuminate the untapped potential of sensory information in refining emotion classification, carving fresh prospects for exploration within the realm of affective computing in NLP.

2. Related work

2.1. Emotion analysis

Over recent years, the domain of pre-trained language models (PLMs) and large language models (LLMs) has witnessed marked advancements. Noteworthy developments include models like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), PaLM (Chowdhery et al., 2022), LLaMA (Touvron et al., 2023) and ChatGPT (OpenAI, 2023). These models, through rigorous pre-training on vast text corpora using self-supervised learning, have the ability to autonomously generate intricate representations. This capability has significantly advanced the field, setting new benchmarks in numerous tasks, notably in sentiment analysis (Devlin et al., 2018; Zhang et al., 2023). An in-depth exploration by Zhang et al. (2023) elucidated the performances of LLMs in sentiment and emotion analysis tasks. The study has highlighted that, while LLMs excel over PLMs in few-shot learning scenarios, PLMs remain superior for more nuanced tasks that demand a deeper understanding of emotions or structured emotional data. Among the discussed models, T5 (Raffel et al., 2020) stands out due to its innovative 'text-to-text' transfer approach, in which every NLP challenge is remodelled as a text-to-text problem. Consequently, T5 frequently sets the state-of-the-arts in emotion analysis when utilized as the base model.

However, despite the considerable improvements made with these PLMs/LLMs, some research gaps remain relatively fulfilled. Present models, although they possess sophisticated neural architectures capable of discerning patterns from immense text datasets, often overlook the intricate nature of emotion—a dynamic interplay of cognitive and physiological responses triggered by various stimuli (Khare et al., 2023). Sensory perceptions, pivotal in shaping these responses, serve as the bedrock upon which our cognitive processes evaluate and generate emotions (Niedenthal and Wood, 2019). Integrating these models with sensory data can potentially elevate their performance, nudging them closer to approaching human-like comprehension. This presents a significant research opportunity: equipping already potent PLMs/LLMs with an element of sensory perception, an aspect they conventionally lack. With our proposed SensoryT5 model, our ambition is to fill this gap by synergizing the strengths of T5 and augmenting it with sensory knowledge, thereby enabling a deeper and more nuanced understanding of emotions.

2.2. Cognition-grounded resources: Sensorimotor norms

In recent years, there is an emergent trend that neuro-cognitive data and computational approaches are synergized in NLP studies. This interdisciplinary synergy unlocks new dimensions in understanding language, sentiment, and emotion, reflecting more accurately the human experience and mental processing. For instance, Long et al. (2019b) improved the attention model for sentiment analysis by incorporating a eye-tracking dataset. Chen et al. (2021) incorporated brain measurement data for modeling word embedding. Wan et al. (2023) demonstrated the superiority of neural networks for metaphor detection by leveraging sensorimotor knowledge. These studies collectively underscore a broader shift in the field towards a more integrated approach to NLP. By weaving in neuro-cognitive data, researchers are equipping computational models with a richer and more intricate understanding of human language and cognition, which are often overlooked by traditional data-driven methods.

Given the intimate connection between emotion and perception as demonstrated in various studies reviewed in the last section, this study assumes that a cognitively and linguistically motivated representation of words in text based on sensorimotor knowledge would improve the performance of computational models for emotion analysis. That is not only because sensory inputs are crucial sources of emotions, but also because emotional responses are part of sensory perceptions for human beings.

Thus, this study utilizes Lynott et al. (2020)'s sensorimotor norms which encompass metrics of sensorimotor strengths (ranging from 0 to 5) of 39,707 concepts spanning six perceptual domains: touch, hearing, smell, taste, vision, and interoception, as well as five action effectors: mouth/throat, hand/arm, foot/leg, head (barring mouth/throat), and torso. To exploit this wealth of data, SensoryT5 is proposed to construct the sensorimotor vectors from these norms and to seamlessly embed them into the T5's decoder mechanism via an auxiliary attention layer. Positioned after the decoders, this sensory-centric attention layer is synergized with the decoder's output, producing an enriched representation brimming with sensory knowledge for words in text. Thus, SensoryT5 is adapted at simultaneously discerning contextual cues and sensory knowledge, allowing for a potent alignment of sensory nuances with contextual intelligence. This integration augments the model's efficacy in the fine-grained emotion classification.

3. Our proposed SensoryT5 model

In this section, we elaborate how our SensoryT5 model incorporates the sensory knowledge into the neural emotion classification framework. Specifically, sensory knowledge is infused into the T5 using an adapter approach built upon attention mechanisms. Moreover, the contextual and sensory information learning branches are amalgamated within a unified loss function to facilitate joint training. The overarching structure is depicted in Figure 1.

3.1. Preliminaries

Despite the relatively large size of the Lancaster Sensorimotor Norms, there are still many out-of-vocabulary words. Following the method proposed by Li et al. (2017), we use a word embedding model to regressively predict the sensory values of unknown words, aiming to obtain sensory values for out-of-vocabulary words.

Inputs and outputs The objective of emotion analysis is to determine and categorize opinions for a piece of texts following a defined label schema. Let D denote a collection of documents for emotion classification. Each document $d \in D$ is first tokenized into a word sequence with maximum length n, then the word embeddings w_i of these sequence are jointly employed to represent the document $d = w_1, w_2, ..., w_i, ..., w_n (i \in 1, 2, ..., n)$.

3.2. The core attention mechanism in T5

The word embeddings of these sequence $d = w_1, w_2, ..., w_i, ..., w_n (i \in 1, 2, ..., n)$ first enters the T5. Each layer of the encoder and decoder has a series of multi-head attention units. The multi-head attention mechanism for the final decoder layer can be represented using the following equation:

$$\begin{aligned} V_d &= \text{MultiHead}(Q_0, K_0, V_0) \\ &= [\text{head}_1, \text{head}_2, ..., \text{head}_i] W_O \end{aligned} \tag{1}$$

Where each head is computed as:

$$\begin{aligned} \text{head}_i &= \text{Attention}(Q_0 W_i^Q, K_0 W_i^K, V_0 W_i^V) \\ &= \text{softmax}\left(\frac{(Q_0 W_i^Q)(K_0 W_i^K)^T}{\sqrt{d_k}}\right) V_0 W_i^V \end{aligned} \tag{2}$$

W_i^Q, W_i^K, and W_i^V are weight matrices that are learned during the training process. They are used to project the input queries (Q), keys (K), and values (V) to different sub-spaces. Q_0, K_0, and V_0 are derived from the output of the penultimate decoder layer. Additionally, following the common practice for text classification with T5, we employ a zero-padding vector as the sole input for the decoder. The result V_d is the output of the T5 decoder, imbued with context-aware attention. Both V_d and K_0 will be utilized in section 3.4 for integration with sensory knowledge.

3.3. Sensory information transformation for T5 integration

We project the Lancaster Sensorimotor Norms into a sensory word vector space. Each word is linked with a six-dimensional vector representing sensory scores across six perceptual modalities (auditory, gustatory, haptic, interoceptive, olfactory and visual dimensions). For a word w, its sensory vector is denoted as $s(w) = [s_1, s_2, ..., s_6]$.

To enable effective integration into the T5-large, we use two linear transformations followed by a

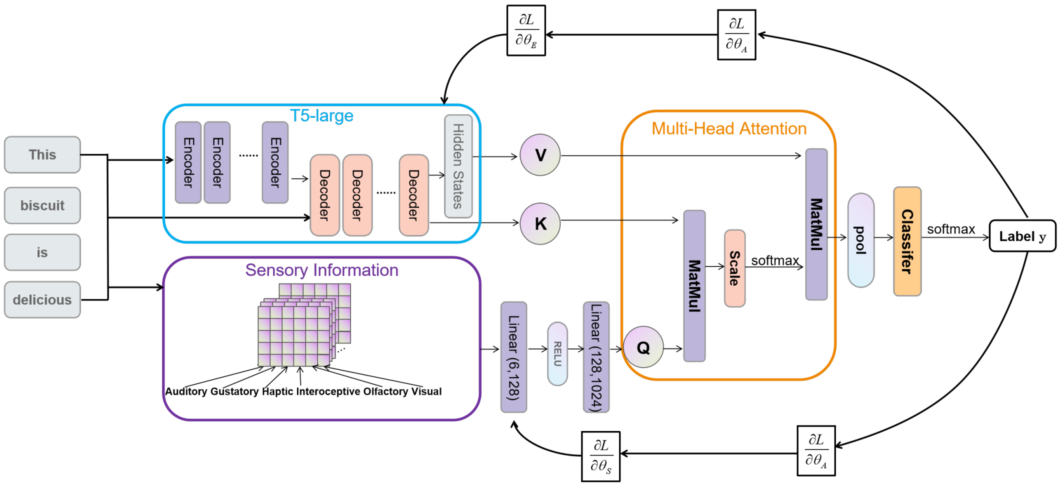


Figure 1: An overview of SensoryT5. Blue box shows a T5 process of deep learning, while purple box describing sensory information is quantified and passed into the T5.

ReLU activation function to map the sensory vectors to the same dimension as the T5-large's word embeddings. Given a T5-large model with an embedding dimension of 1024 , the transformation process can be formally described as:

$$\mathbf{h}_1 = \text{ReLU}(\mathbf{W}_1 s(w) + \mathbf{b}_1) \tag{3}$$

$$s'(w) = \mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2 \tag{4}$$

where $\mathbf{W}_1 : \mathbf{R}^6 \to \mathbf{R}^{128}$ and $\mathbf{W}_2 : \mathbf{R}^{128} \to \mathbf{R}^{1024}$ are two linear transformation matrices and $\mathbf{b}_1$, $\mathbf{b}_2$ are the respective bias terms.The shapes of the two weight matrices W_1 and W_2 are respectively $(6, 128)$ and $(128, 1024)$. The output h_1 of the first linear layer is a vector of shape $(1, 128)$, and the output $s'(w)$ of the second linear layer is a vector of shape $(1, 1024)$. After the transformation, the sensory vector $s'(w)$ is projected into the same semantic space as the features generated by T5-large. The output vector $s'(w)$, with V_d and K_d from the T5, will be applied in section 3.4 for infusing sensory knowledge into T5.

3.4. Sensory attention mechanism in SensoryT5

The sensory vector $s'(w)$ generated by the sensory vector transformation is used as the queries in the attention mechanism of the sensory adapter, substituting the query vector Q in the T5. The sensory adapter performs the attention calculation as follows:

$$\begin{aligned} A_d &= \text{MultiHead}(s'(w), K_0, V_d) \\ &= [a_1, a_2, ..., a_i] W_d \end{aligned} \tag{5}$$

where each head is computed as:

$$\begin{aligned} a_i &= \text{Attention}(s'(w) W_i^Q, K_0 W_i^K, V_d W_i^V) \\ &= \text{Softmax}\left(\frac{(s'(w) W_i^Q)(K_0 W_i^K)^T}{\sqrt{d_k}} \right) V_d W_i^V \end{aligned} \tag{6}$$

Once the output $A_d = a_1, a_2, ..., a_n$ of the sensory adapter is obtained, we apply dropout and pooling operations to form a final representation P_d, which is then used as the input to the classification layer.

$$P_d = \text{Dropout}(\text{Pool}(A_d)) \tag{7}$$

The pooled representation P_d is then fed into the classifier of the T5.

$$C_d = \text{Softmax}(Linear(\text{Dropout}(P_d))) \tag{8}$$

C_d is a probability distribution vector. The class with the highest probability is selected as the predicted label, denoted as y.

The first step of the back-propagation process involves computing the gradient of the loss function with respect to the parameters of sensory attention adapter. Θ_A represents the parameters of the sensory attention layer, and A_d represents the output of the sensory T5. The computed gradient is used to update the parameters of the attention layer, enhancing its capacity to integrate sensory information into the T5 model. This is computed as follows:

$$\frac{\partial \mathcal{L}}{\partial \Theta_\mathrm{A}} = \frac{\partial \mathcal{L}}{\partial A_d} \cdot \frac{\partial A_d}{\partial \Theta_\mathrm{A}} \tag{9}$$

After the gradients for the sensory attention mechanism have been computed, we then compute the gradients for the parameters of the final layer of the T5, denoted as Θ_E.

$$\frac{\partial \mathcal{L}}{\partial \Theta_{\mathrm{E}}} = \frac{\partial \mathcal{L}}{\partial V_d} \cdot \frac{\partial V_d}{\partial \Theta_{\mathrm{E}}} \tag{10}$$

Finally, the gradients for the sensory information transformation, denoted as Θ_{S}, are computed as follows:

$$\frac{\partial \mathcal{L}}{\partial \Theta_{\mathrm{S}}} = \frac{\partial \mathcal{L}}{\partial s'(w)} \cdot \frac{\partial s'(w)}{\partial \Theta_{\mathrm{S}}} \tag{11}$$

Here, Θ_{S} represents the parameters of the sensory information transformation component, which includes the weights and biases of the two linear layers, and $s'(w)$ represents the output of this component. The calculated gradient is used to update the parameters of the sensory information transformation to improve its ability to capture and model sensory information. Through these calculations, we are able to update the parameters of the sensory attention mechanism, the T5, and the sensory information transformation component.

4. Experimental evaluation

4.1. Datasets

We have selected four benchmark datasets of varying sizes to encompass a variety of classification tasks: Empathetic Dialogues (ED) (Rashkin et al., 2019), GoEmotions (GE) (Demszky et al., 2020a), ISEAR (Scherer and Wallbott, 1994) and EmoInt (Mohammad and Bravo-Marquez, 2017). For the GE dataset, we exclusively utilize samples with a single label and omit those that are neutral to maintain an equitable comparison with prior studies (Suresh and Ong, 2021a; Chen et al., 2023). Table 1 presents a summary of key statistics for these datasets. Our evaluation utilizes two widely recognized performance metrics: accuracy and the F1 score, in line with state-of-the-art studies.

Dataset	N_{train}	N_{test}	L	C
ED	19,533	2,547	18	32
GE	23,485	2,984	12	27
ISEAR	4,599	1,534	22	7
EmoInt	3,612	3,141	16	4

Table 1: Statistics of the four benchmark datasets. In the table, "N_{train}" and "N_{test}" respectively represent the number of instances in the training and testing sets. "L" stands for the average text length within the dataset, and "C" indicates the number of classes/categories.

4.2. Sensory knowledge

Before conducting the emotion analysis experiments, we conducted a preliminary analysis of our sensory lexicons from the perspective of sensory perception value distribution. Figure 2 displays histograms of the six sensory measures across all words within our model. Notably, the distributions for these measures are quite unbalanced. Gustatory and olfactory measures predominantly demonstrate a left-skewed distribution, with most values ranging between 0 and 1. This suggests that these two sensory perceptions are less frequently represented in the textual context. Thus, it might be challenging to represent gustatory and olfactory perceptions from text.

In contrast, auditory and visual measures show a relatively uniform distribution. The auditory measure is evenly distributed between 0 and 2.5, while the visual measure ranges between 2 and 4.5. These distributions indicate a higher sensitivity of auditory and visual knowledge to textual information, which suggests that auditory and visual senses may play a significant role within sensory models.

Lastly, haptic and interoceptive measures exhibit similar trends, declining from about 2500 to 0 as the values increase from 0 to 5. The decline in the presence of haptic and interoceptive knowledge across the general textual context might suggest that they are less informative sensory dimensions in the majority of cases.

As discussed in section 3.1, the Lancaster Sensorimotor Norms dataset is subject to size limitations, resulting in a significant number of unknown words for which corresponding sensory values are unavailable. To address this challenge, we adopted the method proposed by Li et al. (2017) for predicting sensory values of unknown words through embedding techniques. In our experiments, we utilized both the T5 embedding and the GloVe embedding (Pennington et al., 2014b) for this prediction task.

To assess the accuracy of our predictions, we randomly selected 10% of the Lancaster Sensorimotor Norms dataset as a validation set and applied the Root Mean Square Error (RMSE) as the evaluation metric. The experimental results are presented in Table 2. The results demonstrate that GloVe outperforms T5 Embedding in predicting each sensory dimension. To preserve the original features of the Lancaster dataset to a minimal extent, we opted for a smaller version of GloVe with 400,000 data points and 200 dimensions. Following augmentation, our sensory vocabulary size reached 407,572[1]

For validating our augmentation, we evaluated the coverage rates of sensory word vectors before and after augmentation across all datasets we employed, as detailed in Table 3. As evident from the augmentation results, the coverage range significantly expands in comparison to the original

[1]The whole dataset of the sensory vocabulary can be accessed at: `https://osf.io/w8yez/?view_only=0e807dfaa5e6433184e452bfebabd01b`.

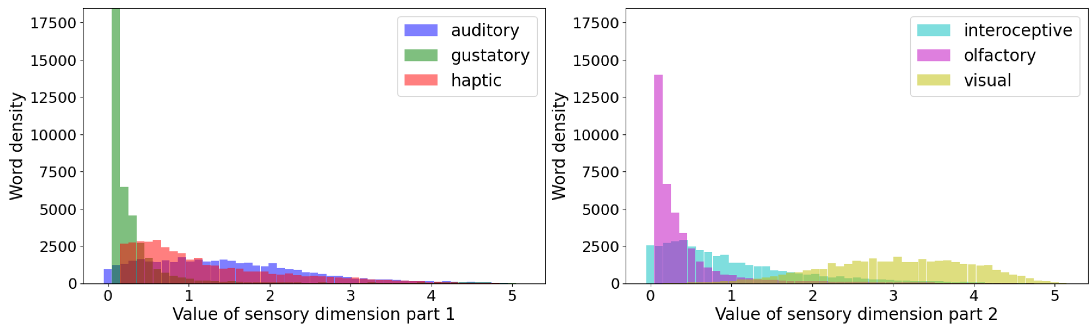


Figure 2: Histograms showing the distribution six sensory values over words. X-axis shows the value in an sensory dimension, while y-axis displays the word density.

Sensory Name	T5 Embedding	GloVe
Auditory	0.949	**0.803**
Gustatory	0.632	**0.534**
Haptic	0.893	**0.698**
Interoceptive	0.831	**0.662**
Olfactory	0.572	**0.501**
Visual	0.842	**0.743**
Total	0.798	**0.665**

Table 2: Comparison of prediction accuracy between T5 Embedding and GloVe techniques on different sensory dimensions, as measured by RMSE values. Lower scores indicate higher accuracy in the prediction of sensory values.

data across all datasets. This underscores the enhanced impact of integrating sensory information into the model on the results.

Datasets	Lancaster %	Exten-Lancaster %
ED	58.23	91.78
GE	46.85	83.91
ISEAR	54.62	78.97
EmoInt	29.65	46.21

Table 3: Word coverage of Lancaster Sensorimotor Norms before and after expansion using regression prediction.

4.3. Experiment settings and Baselines

We compare the proposed SensoryT5 primarily with two group of strong baselines:

PLMs. We compared against BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and T5 (Raffel et al., 2020). The advent of PLMs has marked a significant improvement across a multitude of tasks in the realm of natural language processing, including text classification. This leap in performance is largely due to the deep and nuanced semantic representations these models extract from the text, facilitating a more profound understanding and interpretation of linguistic content.

Label Embedding-aware models. Suresh and Ong (2021a) introduced a concept called label-aware contrastive loss (LCL). This technique uniquely assigns varying weights to each negative sample. Importantly, pairs that are more easily confounded have a higher impact on the objective function, enhancing outcomes in fine-grained text classification scenarios. Chen et al. (2023) proposed HypEmo, a framework enhancing fine-grained emotion classification by utilizing hyperbolic space for label embedding. This model integrates hyperbolic and Euclidean geometries to discern subtle nuances among labels effectively.

These two models, LCL and HypEmo, stand as the most potent in the realm of fine-grained emotion classification, delivering unparalleled results due to their innovative handling of nuanced label distinctions and hierarchical intricacies.

Implementation Details. During training, we applied the Adam optimizer in Euclidean space. We set the learning rate at a consistent 10^{-4}, maintaining a balance between rapid adaptation and the stability of learning, reducing the likelihood of oscillation or divergence.

4.4. Baseline comparison

To demonstrate the effectiveness of SensoryT5, we embarked on a comprehensive set of comparative experiments, analyzing its performance in emotion classification tasks. The comparison is shown in Table 4. Firstly, we compare SensoryT5 with PLMs. SensoryT5 registers an impressive enhancement over T5's performance, the best of the PLM contenders. For instance, SensoryT5 exhibits an increase in accuracy by 0.9% for Empathetic Dialogues and 1.3% for GoEmotions, showcasing its finesse in handling diverse emotional contexts. This upward trend continues with ISEAR and EmoInt datasets, where SensoryT5 improves by 0.9% and 1.2%, respectively, over T5.

Secondly, we compare with label-aware system.

	Empathetic Dialogue		GoEmotions		ISEAR		EmoInt	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
$\star$BERT$_{large}$	0.557	0.551	0.642	0.637	0.677	0.679	0.848	0.848
$\star$RoBERTa$_{large}$	0.596	0.590	0.652	0.644	0.723	0.720	0.865	0.865
$\star$XLNet$_{large}$	0.599	0.592	0.641	0.568	0.711	0.711	0.845	0.845
$\star$T5$_{large}$	0.609	0.604	0.661	0.657	0.717	0.717	0.863	0.863
†LCL	0.601	0.591	0.655	0.648	0.724	0.724	0.866	0.866
§HypEmo	0.596	0.610	0.654	0.663	0.707	0.712	0.846	0.846
$\star$SensoryT5	**0.618**	**0.615**	**0.674**	**0.670**	**0.726**	**0.724**	**0.875**	**0.875**

Table 4: Evaluation on fine-grained emotion classification, the result with the best performance are highlighted in bold. Data marked with †are from (Suresh and Ong, 2021b), §from (Chen et al., 2023), and $\star$represents our own results. Note: In §, results from missing datasets (ISEAR and EmoInt) were supplemented by our experiments.

These two models, LCL and HypEmo, stand as the most potent in the realm of fine-grained emotion classification. LCL outperforms T5 in the ISEAR and EmoInt datasets, while the other datasets under the label-aware system category do not compete favorably with T5. This comparative analysis is critical, considering that LCL utilizes a synonym substitution technique to effectively double its dataset size. Such an expansion contributes significantly to its enhanced performance metrics. In our experiments, we strictly adhered to using original samples without resorting to any form of data augmentation techniques. Despite this, SensoryT5 surpasses LCL by 0.2% and 0.9% in accuracy on the ISEAR and EmoInt datasets, respectively. This margin of improvement, although seemingly nominal, is quite significant in the context of these tasks. It underscores the efficacy of our proposed method of infusing sensory perceptions into the model.

In summary, compared to previous studies, we have achieved superior results without the necessity for additional data, marking the current pinnacle in this field. This accomplishment underscores the effectiveness of SensoryT5.

4.5. Ablation studies

In our efforts to understand the contributions of different components within the SensoryT5 model, we conducted ablation studies, a critical methodological step in assessing the impact of our novel sensory integration. These studies were also carried out on four datasets. The ablation tests were structured around three primary configurations:

SensoryT5: Our complete model infusing sensory information.

Random SensoryT5: A variant of our model where the sensory values were substituted with random numbers ranging from 0 to 5, maintaining the same distribution of sensory scores but eliminating their meaningful association with the data.

T5 (None): The baseline model without any sensory information, representing the standard PLM approach in fine-grained emotion classification tasks.

The result is shown in Figure 3. While the SensoryT5 model exhibited the highest performance in terms of accuracy across all datasets, the Random SensoryT5 configuration yielded lower results than even the T5. This decrement in performance was especially pronounced on the more complex datasets, Empathetic Dialogues and GoEmotions.

The degradation in performance with random sensory values underscores the importance of meaningful sensory integration. It is not merely the presence of additional numerical data that enhances the SensoryT5 model's performance, but rather the contextually relevant and accurately associated sensory information that it brings to the emotion classification task.

Furthermore, the fact that the Random SensoryT5 underperformed compared to the T5 indicates that arbitrarily added sensory information could introduce noise into the model, disrupting its ability to correctly interpret and classify emotional content. This revelation is significant, affirming that the strategic integration of sensory data is crucial, and haphazard integration could be counterproductive.

In summary, these ablation studies have confirmed the value of our sensory information layer, as evidenced by the performance drop when this layer is randomized or removed. This reinforces our assertion that the SensoryT5's strength lies in its ability to simulate a more human-like understanding of textual data, resonating with how humans perceive emotions through a sensory lens.

4.6. Case study

We conducted a focused case study on the SensoryT5 model using a sentence from the Empathetic Dialogues dataset: "I get so mad when I see or hear about kids getting bullied..." In Figure 4, attention heatmaps display the model's focus during processing. The SensoryT5 heatmap shows the aggregate attention for each token in the sensory layer, while the T5 section compiles attention weights across

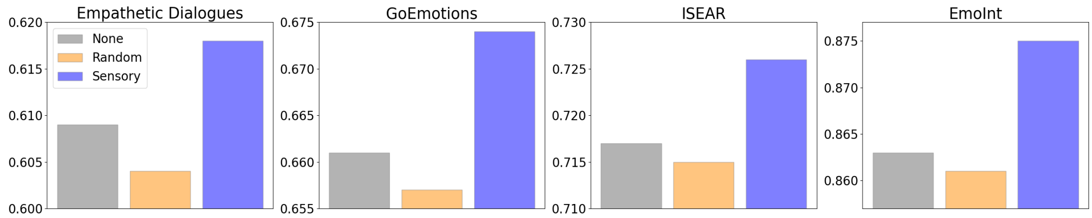


Figure 3: Ablation Study Results. Performance of T5 (None), Random SensoryT5 (with sensory values randomly assigned), and SensoryT5 across four datasets, evaluated using accuracy as the metric.

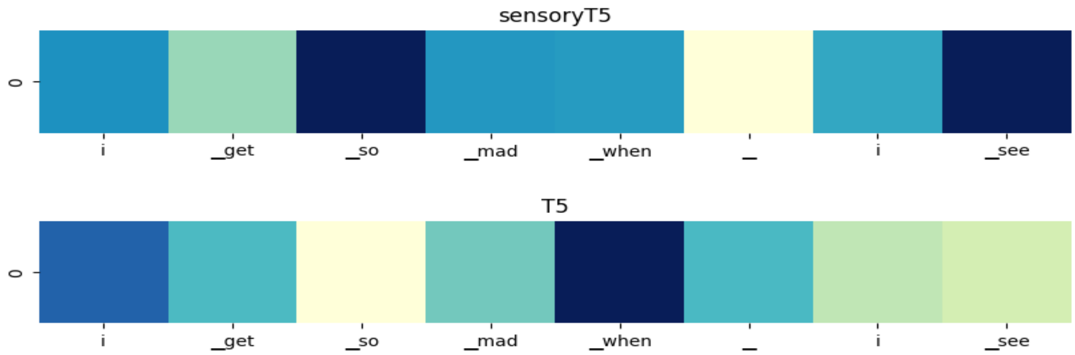


Figure 4: The heat values of the final sensory layer in SensoryT5 and the encoder layer in T5 for the sentence 'I get so mad when I see or hear about kids getting bullied...' sourced from the Empathetic Dialogues training dataset.

all encoder layers, subsequently averaging them to reveal the model's overall focus. The SensoryT5 model exhibited intensified attention on the emotionally significant phrase "so mad," highlighting its ability to detect crucial emotional nuances. In contrast, the standard T5's attention was more distributed, less focused on the emotional pivot. This micro-level analysis reveals SensoryT5's superior capability in recognizing emotional cues. Such insights substantiate the efficacy of integrating sensory awareness into language models for improved emotional discernment.

In summary, our extensive evaluations and comparative studies highlight the superior performance of SensoryT5 over other PLMs based emotion classification models, including the T5. When benchmarked against the state-of-the-art methods, SensoryT5 notably surpassed them, establishing a new standard in the field. Further, our ablation studies convincingly demonstrate that the effectiveness of SensoryT5 is attributed more to its integration of sensory perception than to structural enhancements. This assertion is corroborated by our detailed case studies, which offer a microscopic view into the instances where SensoryT5's unique capabilities are distinctly evident. Collectively, these findings underscore a breakthrough performance of SensoryT5 in the realm of fine-grained emotion classification. Importantly, it signifies a successful adaptation within the shift towards incorporating neuro-cognitive data in NLP, validating the premise that a deeper convergence between sensory data and language modeling leads to a more profound understanding of emotional nuances.

5. Conclusion

In this paper, we propose the SensoryT5 model designed for the fine-grained emotion classification. This framework harnesses sensory knowledge, aiming to boost the prowess of transformers in pinpointing nuanced emotional subtleties. By integrating sensory knowledge into T5 through attention mechanisms, the model concurrently evaluates sensory cues alongside contextual hallmarks. Crucially, SensoryT5 exhibits exceptional adaptability and precision, making it a formidable tool for tasks in Fine-grained Emotion Classification, including configurations like 32-class, 27-class, 7-class, and 4-class delineations. Moreover, SensoryT5 serves as a conduit between sensory perception and emotional understanding, embodying the recent paradigm shift in NLP towards a more neuro-cognitive approach. It acknowledges and capitalizes on the intrinsic relationship between our sensory experiences and our emotional responses, a connection well-documented in neuro-cognitive science but often under-explored in computational fields. By interpreting sensory lexicon through advanced representation learning, SensoryT5 de-codes the implicit emotional undertones conveyed, mirroring the human ability to associate sensory experiences with specific emotional states. In recognizing the entwined nature of cognition, sensation and emotive expression, SensoryT5 not only contributes to but also encourages the continuation of interdisciplinary research efforts. It stands as testament to the potential of a more nuanced and integrative approach in NLP, where understanding language transcends the boundaries of words and grammar, delving into the very experiences and perceptions that shape human emotionality.

Limitations

In our work, we utilized GloVe and T5 embeddings to predict sensory values for unknown words using a regression method. This approach learns only from static values. To derive static T5 embeddings, we passed all tokens sequentially through the T5 embedding layer, obtaining a static embedding for each token. This process, however, leads to a limitation: it compromises the original dynamic context-embedding capabilities of T5. In T5 embeddings, different embeddings are obtained based on the different contexts. We intended to learn from these transformer embeddings and then predict. Additionally, when compared to current state-of-the-art models in emotion classification, such as the label embedding-aware HypEmo and LCL, SensoryT5 exhibits certain inadequacies, particularly in terms of interpretability. Both HypEmo and LCL not only surpass SensoryT5 in explaining their decision-making processes but also do so with fewer parameters. These models, by leveraging sophisticated label-aware embedding strategies, provide insights into the nuanced relationships and hierarchies among labels, something that SensoryT5, with its reliance on static values, struggles to achieve. This gap highlights a significant area for improvement in SensoryT5, suggesting the need for an advanced approach that maintains the richness of context-sensitive embeddings while enhancing the model's overall interpretability and efficiency.

Acknowledgements

This work is supported by the Alan Turning Institute/DSO grant: Improving multimodality misinformation detection with affective analysis. Yunfei Long, and Yuhan Xia acknowledge the financial support of the School of Computer science and Electrical Engineering, University of Essex.

6. Bibliographical References

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Mukul Bhalla and Dennis R Proffitt. 1999. Visual–motor recalibration in geographical slant perception. *Journal of experimental psychology: Human perception and performance*, 25(4):1076.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chih-Yao Chen, Tun-Min Hung, Yi-Li Hsu, and Lun-Wei Ku. 2023. Label-aware hyperbolic embeddings for fine-grained emotion classification. *arXiv preprint arXiv:2306.14822*.

I-Hsuan Chen, Qingqing Zhao, Yunfei Long, Qin Lu, and Chu-Ren Huang. 2019. Mandarin chinese modality exclusivity norms. *PloS one*, 14(2):e0211336.

Xin Chen, Zhen Hai, Suge Wang, Deyu Li, Chao Wang, and Huanbo Luan. 2021. Metaphor identification: A contextual inconsistency based neural sequence labeling approach. *Neurocomputing*, 428:268–279.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Chloe Clavel and Zoraida Callejas. 2015. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing*, 7(1):74–93.

Edward Collins, Nikolai Rozanov, and Bingbing Zhang. 2018. Evolutionary data measures: Understanding the difficulty of text classification tasks. *arXiv preprint arXiv:1811.01910*.

Louise Connell, Dermot Lynott, and Briony Banks. 2018. Interoception: the forgotten modality in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752):20170143.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020a. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020b. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

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*.

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.

Ruihai Dong, Michael P O'Mahony, Markus Schaal, Kevin McCarthy, and Barry Smyth. 2013. Sentimental product recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 411–414.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Lynn Fainsilber and Andrew Ortony. 1987. Metaphorical uses of language in the expression of emotions. *Metaphor and Symbol*, 2(4):239–250.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Minghui Huang, Haoran Xie, Yanghui Rao, Yuwei Liu, Leonard KM Poon, and Fu Lee Wang. 2020. Lexicon-based sentiment convolutional neural networks for online review analysis. *IEEE Transactions on Affective Computing*, 13(3):1337–1348.

Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 720–728.

Xiaotong Jiang, Qingqing Zhao, Yunfei Long, and Zhongqing Wang. 2022. Chinese synesthesia detection: New dataset and models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3877–3887.

Patrik N Juslin and Daniel Västfjäll. 2008. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, 31(5):559–575.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2019. Sentilare: Sentiment-aware language representation learning with linguistic knowledge. *arXiv preprint arXiv:1911.02493*.

Smith K Khare, Victoria Blanes-Vidal, Esmaeil S Nadimi, and U Rajendra Acharya. 2023. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, page 102019.

Zoltán Kövecses. 2019. Perception and metaphor. *Perception metaphors*, 19(327):10–1075.

George Lakoff and Mark Johnson. 1980. Metaphors we live by. *University of Chicago, Chicago, IL*.

Sophia Yat Mei Lee. 2018. Figurative language in emotion expressions. In *Chinese Lexical Semantics: 18th Workshop, CLSW 2017, Leshan, China, May 18–20, 2017, Revised Selected Papers 18*, pages 408–419. Springer.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, 8(4):443–456.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. *arXiv preprint arXiv:2111.02194*.

Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. 2022. Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7124–7135, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yunfei Long, Rong Xiang, Qin Lu, Chu-Ren Huang, and Minglei Li. 2019a. Improving attention model based on cognition grounded data for sentiment analysis. *IEEE transactions on affective computing*, 12(4):900–912.

Yunfei Long et al. 2019b. A study on using personal profiles for a biased reader emotion prediction model.

Qiang Lu, Xia Sun, Yunfei Long, Zhizezhang Gao, Jun Feng, and Tao Sun. 2023. Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. *IEEE Transactions on Neural Networks and Learning Systems*.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52:1271–1291.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Donatas Meškelė and Flavius Frasincar. 2020. Aldonar: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. *Information Processing & Management*, 57(3):102211.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Nadine Müller, Arne Nagels, and Christina Kauschke. 2021. Metaphorical expressions originating from human senses: Psycholinguistic and affective norms for german metaphors for internal state terms (mist database). *Behavior Research Methods*, pages 1–13.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81.

Paula M Niedenthal and Adrienne Wood. 2019. Does emotion influence visual perception? depends on how you look at it. *Cognition and Emotion*, 33(1):77–84.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014a. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014b. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Dennis R Proffitt, Mukul Bhalla, Rich Gossweiler, and Jonathan Midgett. 1995. Perceiving geographical slant. *Psychonomic bulletin & review*, 2:409–428.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Cedar R Riener, Jeanine K Stefanucci, Dennis R Proffitt, and Gerald Clore. 2011. An effect of mood on the perception of geographical slant. *Cognition and Emotion*, 25(1):174–182.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Goran Šimić, Mladenka Tkalčić, Vana Vukić, Damir Mulc, Ena Španić, Marina Šagud, Francisco E Olucha-Bordonau, Mario Vukšić, and Patrick R. Hof. 2021. Understanding emotions: Origins and roles of the amygdala. *Biomolecules*, 11(6):823.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 151–161.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Varsha Suresh and Desmond Ong. 2021a. 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.

Varsha Suresh and Desmond C Ong. 2021b. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. *arXiv preprint arXiv:2109.05427*.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Mingyu Wan, Qi Su, Kathleen Ahrens, and Chu-Ren Huang. 2023. Perceptional and actional enrichment for metaphor detection with sensorimotor norms. *Natural Language Engineering*, pages 1–29.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Rong Xiang, Jing Li, Mingyu Wan, Jinghang Gu, Qin Lu, Wenjie Li, and Chu-Ren Huang. 2021. Affective awareness in neural sentiment analysis. *Knowledge-Based Systems*, 226:107137.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Takashi Yamamoto. 2008. Central mechanisms of taste: Cognition, emotion and taste-elicited behaviors. *Japanese Dental Science Review*, 44(2):91–99.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Sruthi Yarkareddy, T Sasikala, and S Santhanalakshmi. 2022. Sentiment analysis of amazon fine food reviews. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1242–1247. IEEE.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. *arXiv preprint arXiv:2005.04114*.

Jonathan R Zadra and Gerald L Clore. 2011. Emotion and perception: The role of affective information. *Wiley interdisciplinary reviews: cognitive science*, 2(6):676–685.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. 2017. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135.

Author Index