# ASRAEL - Acquisition of Semantic RelAtions bEtween Latin nouns

Konstantin Schulz

Humboldt-Universität zu Berlin

January 17, 2022

# Contents

# List of Figures

# List of Tables

# Glossary

**CVS** continuous vector space. 4, 5, 15, 16

**DSM** distributional semantic model. 3, 5

**PMI** Pointwise Mutual Information. 3

**POS** part of speech. 5

**TTR** type-token ratio. ii, 19, 20

**WSD** word sense disambiguation. 6

**WSI** word sense induction. 5, 6

# abstract

XXXabstractXXX

# Chapter 1

# Introduction

This dissertation deals with different ways of detecting semantic relations in text corpora. Its focus is more on methodology than on analytical results. Some important restrictions are necessitated by the precise research question: In a corpus of Latin texts, how can we automatically extract semantic relations that are relevant for language learning? I will concentrate only on a subset of the existing phenomena, namely hypernymy and hyponymy. This selection is motivated by the objective of utilizing the obtained insights as the basis for exercises in the process of learning historical languages. Thus, the detection of said relations is not an end in itself, but is to be viewed in the context of how humans learn languages.

## 1.1 Semantic relations in language learning

Specifically, recent studies suggest that an improved integration of hyponymy into the learning process can lead to a higher level of language skills (Taslim 2014, p. 196). Roughly the same procedure is applicable also in the context of artificial intelligence, which highlights the general importance of semantic relations in any kind of language learning effort (Carlson et al. 2010, p. 1306). The automated extraction of semantic relations from texts can be used for constructing ontologies in specific domains, which is important for language learning because many textbooks are heavily structured according to a range different topics (Punuru et al. 2012, 192f.).

Since there is a parallel between semantic relations (or ontologies in general) and the way that humans think, it seems reasonable to exploit that connection for learning purposes. This is particularly important for advanced learners' lexical acquisition (S. A. Crossley et al. 2010, p. 56). Now, one might be tempted to conclude that predefined lists of basic vocabulary and more advanced semantic relations would be a sufficient help to language teachers. However, static materials und human supervision will probably fail to account for the entire processual complexity of lexical acquisition (S. A. Crossley et al. 2010, p. 71).

The involvement of artificial intelligence in the modeling of human knowledge about vocabulary faces various problems, including the progression of learning over time (S. A. Crossley et al. 2010, p. 70).

In language learning contexts, this problem extends also to the morphological domain, where it can prove difficult for students to exploit their knowledge about affixation for the acquisition of semantic relations (Gardner 2007, p. 248). However, "inflected and derived forms" are exactly what we encounter in the curricular targets for Latin courses, i.e. literary texts. Most textbooks take care of the problem by using hand-crafted Latin texts instead of authentic ones. This leads to a highly problematic transition from textbook to literature, which is already famous among German teachers of Latin (Schibel 2013, p. 115). The challenge, then, is to not only extract semantic relations from a given corpus, but also to group them according to general principles of progression that are tightly coupled with a domain-specific theory of learning.

# Chapter 2

# Polysemy

## 2.1 Word Sense Induction

In order to reach that point, we first need to define the basic object to be acquired, i.e. semantic relations. Unfortunately, this is not a precisely defined research object per se because it strongly depends on our previous knowledge about the possible word senses (Ayşe et al. 2011, p. 12). Consequently, the sense of a word has to be defined before we can start to learn more abstract representations like hyponymy (Bartunov et al. 2016, p. 137). Further obstacles may occur in the case of specific relations like synonymy, where a good operationalization is needed to retrieve the relevant instances from a corpus of text data (Divjak et al. 2009, p. 274). In my study, I will focus heavily on the approach of distributional semantics, describing specific semantic relations between words in terms of their co-occurrences, i.e. typical contexts in a corpus (Harris 1954, p. 162; Firth 1957, p. 30). This approach has traditionally suffered from several limitations. In some cases (e.g. synonymy), building unified representations for every type in a corpus can prove highly problematic (Karan et al. 2012, p. 114). However, this is not universally true anymore as there are DSMs that consistently distinguish between various contexts (see below). Other problems of distributional semantics are tied to fine-grained distinctions between specific semantic relations (Karan et al. 2012, p. 115).

Co-occurrence counts can be a good basis for the detection of semantic relations in general, but they are rarely sufficient for a thorough classification into several subcategories. Nonetheless, previous research has been carried out (more or less successfully) using, e.g., association measures to detect synonymy (Hagiwara et al. 2009, p. 566). One of the critical parts in that regard is the term 'word pair', which introduces the linguistically ill-defined concept of 'word' into an otherwise seemingly precise mathematical formula. In the following, we will start by using the term 'word' simply as 'token in a corpus' and later proceed to a more advanced definition. This progression seems reasonable given the fact that the usage of token-based PMI for the research purposes of distributional

semantics often leads to severely skewed results (Herbelot and Ganesalingam 2013, p. 444). Therefore, corpora may have to be tokenized and segmented using more complex separator rules than 'whitespace and punctuation' before they can serve as a basis for distributional analyses (Mikolov, Sutskever, et al. 2013, p. 6). Besides, polysemy may also become a problem for resolving other issues like synonymy, e.g. if two words are synonymous only in one of their possibly numerous senses (Herbelot and Ganesalingam 2013, p. 444).

In addition to polysemy itself, related concepts like homonymy can also become obstacles with regard to the aim of defining semantic relations precisely (Gardner 2007, p. 251). Distributional semantics provides a valuable framework to break down the multiple aspects of meaning for homonymous or polysemous words into smaller subsets, based on the combination with, e.g., adjectival modifiers (Boleda et al. 2013, p. 42). Therefore, the integration of specific contextual elements into the analysis of hyponymy between Latin nouns seems to be a reasonable basis for distinguishing subtle semantic nuances more consistently. As a consequence, polysemy will be treated in this study not just as one of the analytical targets, but also as a methodological aspect that needs to be addressed before other kinds of analyses can be performed. A useful operationalization of polysemy in a distributional context is to look at the collocational diversity (Hamilton et al. 2016, p. 8). This approach brings along the problem of sampling, which is a general one for the whole of corpus linguistics: If we want to model a specific language like Latin as a whole but only use a tiny subset of the available language data for our corpus, how can we expect our results to be applicable to other Latin texts? The key point here is the purpose: For this study, our ultimate goal is not to model Latin nouns per se, but only with regard to the rather well-defined application context of language learning. Therefore, the explanatory power of our linguistic model does not need to be as comprehensive, and it will be restricted to predefined parameters. These parameters may in turn be described as, e.g., the objective of maximizing a student's learning success by providing structured information about semantic relations.

## 2.2 Distributional Semantic Models for Word Senses

Historically, linguistic features of words have often been represented in CVSs (Mikolov, Chen, et al. 2013; Dai et al. 2015; Akbik et al. 2018) that can be seen as statistical models of word distributions in natural language (Bengio et al. 2003, p. 1140). Such models map single elements of an utterance (e.g. characters, tokens, types, phrases etc.) to vectors in an n-dimensional CVS. Depending on the dimensionality, various kinds of information will be integrated into the model. For instance, lexically related words (e.g. *amor* 'love' in PROIEL 86055.22[1] and *amare* 'to love' in PROIEL 86334.11) usually correspond to vectors whose

[1]For details on the PROIEL corpus, see Haug et al. (2008). Note that, in this study, we only consider the Latin part of the corpus. Dot-separated unique identifiers will be used to refer to segments and tokens in the PROIEL corpus. This is often more precise than the conventional citation model for ancient works (in this case: Cic. off. 1.12) and does not require knowledge about abbreviations of author names or work titles. On the downside, it is

distance in the CVS is rather small. Since most of these models try to predict words in a corpus given their surrounding context (Mikolov, Chen, et al. 2013, p. 5), those words that co-occur frequently will also be mapped to similar vectors in the CVS. This basic interaction helps us to distinguish multiple senses of a polysemous word empirically: If each word sense is associated with different contexts and, thus, different co-occurring words, its calculated vector will be different from those of other word senses. Therefore, a very simple evaluation for DSMs with regard to WSI looks like this:

- Given a pair of contexts that
- both contain the same word,
- classify whether that target word is used in the same sense or not.

Formalizing this into a dataset structure, we may provide a plain text file where each each line contains one example for classification. Since Latin is a highly inflectional language, morphological features of the targeted word can differ a lot. Therefore, the actual word forms have to be indicated separately for each context in the example. If they occur multiple times in a context, the dataset should specify which occurrence of the word is to be evaluated. Finally, a binary value (such as '0 / 1' or 'true / false') indicates the ground truth for this classification example, i.e. whether the word senses are the same or not. This is the most critical part: 'Ground truth' is a dangerous assumption for linguistic phenomena because it suggests a consistent, unchanging goal to be achieved (Lee et al. 2012, p. 129). This desire stems from the formalization that is deemed necessary to systematically evaluate computational language models (Fischer et al. 2010, p. 4). Usually, such gold standards depend on human expertise, i.e. people that are highly acquainted with the target domain (Uzuner et al. 2010, p. 519). In the case of historical languages like Latin, where specialized software and datasets are scarce (Karakanta et al. 2018, p. 168; McGillivray 2013, p. 3), researchers often have to resort to digitizing analog resources or creating entirely new ones. This kind of curation process can and should rely on successful patterns from datasets for modern languages. Thus, it is reasonable to have a short look at existing comparable resources for evaluation.

The SemCor (Mihalcea 2008) and MASC-News (Moro et al. 2014, p. 4218) datasets are POS-tagged English corpora where word senses have been annotated using WordNet (Tengi 1998). In that sense, such data is both empirical (because it contains authentic historical utterances) and intuition-based (because it uses an external knowledge base for defining the word senses, see below). There have been attempts to unify (Navigli et al. 2010, p. 216) the various knowledge bases that researchers employed to annotate word senses. Promising results were achieved by discussing discrepancies in the face of text corpora, i.e. actual text passages from authentic sources (Baker et al. 2009, p. 127). This development highlights the importance of usage-based linguistics

corpus-specific and does not offer immediate information about author, work and structural position of the text passage.

and points towards the need for a more rigorous, bottom-up formalization of word senses. Further approaches to creating datasets for WSD include the usage of dictionaries (other than WordNet) as lexical resources (Mihalcea et al. 2004, p. 25; Jimeno-Yepes et al. 2011, p. 2; D. Yuan et al. 2016, p. 6) and the usage of glosses as representations of word senses (F. Luo et al. 2018, p. 8). In contrast, WSI usually does not rely on external sense inventories, but tries to infer word senses in an unsupervised, empirical manner only from the text itself. State-of-the-art models cluster distributions of vector representations for possible context-dependent substitutes of a word and, thus, are able to detect distinctions between word senses that had not previously been considered in intuition-based sense inventories (Amrami et al. 2019, pp. 1–4). Predecessors to this approach directly used a word's surrounding context to construct sense representations for clustering and identified substantial quality criteria for the evaluation of WSI models (Huang et al. 2012, pp. 876–878):

- human judgment on word sense similarity
- variation in patterns of meaning (homonymy, polysemy etc.)
- variation in parts of speech (esp. nouns, verbs, adjectives)

We can certainly add that, in addition to variation, balancing is a decisive factor for the training data (Liu et al. 2011, p. 620; Wei et al. 2013, p. 9; Bone et al. 2015, p. 1127): If we are interested in rare linguistic phenomena, such as subtle semantic differences between almost synonymous words, we need to make sure that they are represented frequently in our training data. The third criterion is most controversial: On the one hand, powerful WSI models need to be general enough to provide sufficient results under different circumstances, e.g. in varying domains and registers. On the other hand, modeling word senses for different parts of speech needs to take into account the diverse influences of lexis and syntax: In some settings, nouns may be represented as sets of their hyponyms (Lewis 2019, p. 640), while transitive verbs can be modeled as triples together with their arguments, e.g. subjects and objects (Grefenstette et al. 2011, p. 9). Such different approaches uncover the intricate nature of formalizing fine-grained semantic distinctions from a distributional perspective. As a consequence, our goal of detecting hyponyms for language learning seems to warrant a modular operationalization with regard to parts of speech.

# Chapter 3

# Nouns

Our choice of nouns as the target for hyponymy detection is motivated by evidence from language learning. It suggests that nouns are usually easier to acquire, especially in most modern languages, where this class of words is particularly important in everyday language use (Garcia-Gamez et al. 2019, p. 9). This is relevant for our research because, nowadays, many learners of Latin have acquired as L1 one of the 'noun-dependent' languages, e.g. German or French (Baïdak et al. 2017, pp. 51–52). Therefore, it makes sense to build on their individual preconditions and focus, at least initially, on those word classes that are also important in their L1. However, Latin has often been described as verb-heavy language (Menge et al. 2009, p. 41). This claim is supported by comparing the noun-to-verb ratio (Stoll et al. 2012, p. 300) of the PROIEL corpus to the Hamburg Dependency Treebank (Borges Völker et al. 2019) and the French Treebank (Seddah et al. 2018; see Table 1).

| | PROIEL | French Treebank | Hamburg Dependency Treebank |
|---|---|---|---|
| Noun-to-Verb Ratio | 0.50 | 0.71 | 0.73 |

Table 1: Noun-to-verb ratio for different treebanks in the Universal Dependencies corpus (Latin, French, German; Nivre et al. 2016).

The lower ratio indicates that Latin authors use fewer nouns, relative to their usage of verbs and compared to authors of modern European languages. This typological difference seems to demand a stronger focus on verbs instead of nouns, at least for Latin. Nonetheless, it has been shown that nouns play a substantial role in the early stages of language acquisition, especially in the understanding of subcategorization frames for verbs (S. Yuan et al. 2012, p. 1383). If this is applicable to historical languages as well, we should rather rely on finding suitable combinations, provided that there is a useful formalization of hyponymy for the different parts of speech. The focus on a single word class (in our case: nouns) is one initial step in that direction.

# Chapter 4

# Hyponymy

After dealing with various issues of polysemy, some of the second-tier semantic relations may have to be redefined when looked at from a distributional perspective, e.g. hyponymy (Herbelot and Ganesalingam 2013, p. 443). In this view, other existing taxonomies that may have acted as models of hyponymy need to be replaced by a definition that reflects how actual usage contexts can limit our selection of possible candidates for hyponymy (Roller et al. 2014, p. 443). Unfortunately, this view may clash with the observations of Herbelot and Ganesalingam (2013):

- Hypernyms can have very specific usage contexts.
- Collocations may lead to unexpected results in the distribution of words.

Similar adjustments have to be made in treating other kinds of semantic relations. Therefore, many researchers apply preprocessing to their corpora before performing the actual analysis on them (Gyllensten et al. 2019, 56f.). However, this increases the danger of circularity, revealing only those results that we were already expecting in the first place. As an alternative, we should look for a model that is robust enough to detect semantic relations despite the confounding factors. The objective should be to fit the model to the base text, not the other way round. For vector space models like Skip-gram (Mikolov, Chen, et al. 2013, p. 5), subsampling and the restriction of strong collocations may be justified from a computational perspective because they enable a faster training of models while more or less preserving the resulting performance in the prediction of word contexts. This very tradeoff, though, may not work out as well for the detection of semantic relations, e.g. if we give a higher priority to the frequency of certain contexts. Such a change of algorithm may particularly be motivated by the distinction of theoretical and empirical scopes of words (Coenen 2013, p. 46).

This also applies to the domains of corpus linguistics and language learning: Given a text corpus that helps learners to acquire a certain knowledge (e.g. all the ancient literature mentioned in an educational curriculum for the Classics),

the theoretical scope of a certain word would be equivalent to the set of all distinct contexts in the corpus in which that word occurs. In contrast, the empirical scope of that same word would be equal to a subset of that theoretical scope, namely those contexts which are already known to the learner (at any given point in time of the learning process). Thus, the theoretical scope will remain static as long as we do not change the corpus, while the empirical one will be dynamic as long as we keep learning. From this point of view, we can also restrict our notion of semantic relations: Hyponymy, for instance, will be tied to those lexical networks that are present in a specific corpus and, for a specific learner, specific instances of hyponomy relations will not exist unless they have already been encountered in the corpus and lead to a relevant learning process. It is therefore our duty to a) detect such instances in the corpus and b) use them to design learning situations that help language learners to construct lexical and semantic knowledge (York et al. 2018, p. 25). In order to accomplish that first part, we need to have clearly defined requirements to qualify a relation as one of hyponymy, e.g. according to a version of the Distributional Inclusion Hypothesis (Roller et al. 2014, p. 443) that also accounts for idiomaticity, burstiness and other kinds of lexical interferences.

## 4.1 WordNet

Lexical resources like the Latin WordNet (Minozzi 2010) may contribute a reference model for such a purpose because it includes a built-in notion of hyponymy. Unfortunately, the WordNet defines such relations based on intuition, not on actual language use (Fellbaum et al. 2012, p. 315). WordNet is therefore static and non-empirical from the perspective of usage-based fields like corpus linguistics. However, it was constructed by alleged experts and we may use their assessment for reflecting our own approach. This kind of triangulation is supported by recent studies, which often explicitly combine knowledge bases and corpus data (Ono et al. 2015, pp. 984–988), thereby integrating both the introspective and the usage-based approaches to semantics. The former can be helpful in two ways: a) by defining a horizon of expectations in language learning contexts, i.e. which information learners are supposed to infer from their previous knowledge in order to understand a linguistic expression in a specific text passage; and b) by modeling semantic knowledge in a machine-readable, consistent (albeit simplistic) manner. The integration of distributional semantics, on the other hand, is particularly motivated by the hypothesis that semantic properties of words are reflected in their contextual surroundings (Gries et al. 2009, p. 59).

By using distributional data from corpora, we also gain the advantages of rigorously empirical research (as opposed to bare intuition and anecdotal evidence): frequency, authenticity, variation and systematic induction (Gries et al. 2009, p. 60). Still, knowledge databases may be useful for languages like Latin where previous research is scarce and, more often than not, heavily intuition-based or anecdotal. In both the introspective and the distributional case, we can increase the amount of provided information by relying on associated properties

of hyponymy relations, e.g. transitivity (Coenen 2013, p. 58). Applied to the Latin WordNet, this may look as follows:

1. *gladius*
2. a cutting or thrusting weapon with a long blade
3. weaponry used in fighting or hunting
4. weapons considered collectively
5. an artifact (or system of artifacts) that is instrumental in accomplishing some end
6. a man-made object
7. a physical (tangible and visible) entity

This information flow can be read from both sides. Beginning at *gladius* (sword), it can be used to simulate the process of reading a text: We encounter a word in the text (e.g. the plural *gladii*) and determine its lemma (*gladius*). From there, we can try to find a corresponding meaning for the surface form, e.g. "a cutting or thrusting weapon with a long blade". This meaning, in the WordNet taxonomy, can be hyponym to other kinds of meaning, e.g. "weaponry used in fighting or hunting". Now, if we know that

- *gladius* is a weapon with a blade and
- weapons with a blade can be used for hunting,

then we are safe to assume that a *gladius* can be used for hunting. The further we go down the list, the more abstract do the meanings get. Nonetheless (or because of that), every layer can remind us of new specific features that we may want to associate with *gladius*: specific motion (2.), specific purpose (3.), hypernym (4.), general purpose (5.), origin and production (6.), material (7.). In some cases, we may even use morphological hints to get faster access to information about hyponymy (Anstatt 2009, p. 913). This is especially true for Latin with its relatively rich morphology: the meaning of *artifex* (artist) may be associated directly to *ars* (art) and *facere* (to make), without necessarily traversing a hierarchical tree-like representation of increasing semantic generalization. Interestingly, this is possible even in cases where the compound is formed from foreign language material and thus cannot be easily paraphrased in the native language (Souillé-Rigaut 2010, p. 33).

How is that important for Latin hyponymy? Consider Greek loanwords: *xylophytum* (a certain kind of plant) cannot be reasonably described in Latin words as a **phytum* that is made of **xylum*. However, Romans may have deduced that the meaning of *xylophytum* is related to (→ hyponymy) *-phyt-*, thereby inferring previous knowledge about plants. Therefore, we have to allow for the possibility that some instances of hyponymy may not be extractable solely on a

distributional basis, but rather on a morphological one. Besides, a morphologically informed model is able to explain why we can immediately see semantic structure in newly coined words (Souillé-Rigaut 2010, p. 43). In this respect, knowledge about specific hyponyms or hyponymy in general can facilitate the language learning process by providing a rather intuitive access to the meaning of single words or phrases. Unfortunately, this intuition sometimes competes with other interpretations of a word's sense which may be equally intuitive and valid (Pons Bordería 2014, p. 126).

Such examples underline the importance of disambiguation (see above), which may not always be possible. Furthermore, additional steps may be necessary even if a specific word sense was already determined, e.g. the inference of common knowledge or common sense (Pinkal 1993, 427f.). In this view, disambiguation should be preceded by a thorough linguistic analysis of the context, which, in terms of distributional semantics, is usually addressed by extracting cooccurrence frequencies from sliding windows over a textual input, be it plain text or a linguistically annotated treebank. However, the integration of external information into this process is not as straightforward and therefore often avoided. In the end, this might not be the worst decision, depending on our definition of semantics: Some models of distributional semantics reduce the spectrum of polysemy to one main aspect of meaning per word. (Faruqui et al. 2016, p. 4).

Approaches of this kind level nuances that may actually be important for our understanding of a single word in its context. As an improvement, other models regard entire sentences for the representation of a word's context-specific meaning, thereby differentiating between various common usages (Peters et al. 2018, 2f.). Modeling the process of language production both in a forward and backward manner at the same time seems reasonable, especially for the Latin language, because the antecedent of, e.g., a relative pronoun can be part of the relative clause, which makes it difficult to understand the structure of such a construction by reading it in just one direction, without referring to the past and future context at the same time:[1]

(PROIEL 53467)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *nam* | *et* | *frument-um* | *ex* | *agr-is* | *cotidie* | *in* | *castr-a* |
| since | also | grain-ACC | from | field-ABL.PL | daily-ADV | into | camp-ACC.PL |
| *confere-ba-t* | | *et* | *qu-ae* | *grav-issim-e* | *adfli-ct-ae* | | |
| convey-IMPF-3SG | | and | which-F.PL | heavy-SUP-ADV | strike-PASS-F.PL | | |
| *era-nt* | *nav-es* | *e-a-rum* | | *materi-a* | *atque* | *ae-re* | *ad* |
| be-3PL | ship-PL | this-F-PL.GEN | | timber-ABL | and | brass-ABL | to |
| *reliqu-a-s* | | *reficie-nd-a-s* | | *ute-bat-ur* | *[...]* | | |
| remaining-F-PL.ACC | | repair-GDV-F-PL.ACC | | use-IMPF-3SG | | | |

for he daily conveyed corn from the country parts into the camp, used the timber and brass of such ships as were most seriously damaged for repairing the rest [...]

[1] The glossing of longer Latin text passages was created according to the Leipzig Glossing Rules.

In this example, *naves* is the antecedent to the relative pronoun *quae*. Its integration into the relative clause is recognizable by its morphological congruency (case, number, gender) with the relative pronoun. In a textbook setting, the sentence would rather look like this: *[...] conferebat et earum navium, quae gravissime adflictae erant, materia atque [...].* Such syntactic variations are probably related to the relatively free word order in the Latin language. Therefore, a language model for Latin should be flexible enough to analyze a given textual sequence by looking at either the left or right context, according to the circumstances of the case (Devlin et al. 2019, p. 8). While this strictly bidirectional, attention-based approach (Vaswani et al. 2017) enhances the ability to explain word meanings from their larger surrounding contexts, it does not address the problem of inference, e.g. relevant external text passages. To analyze this problem empirically, we will set up a toy language model in order to inspect difficult cases.

# Chapter 5

# Method

## 5.1 Experimental Setup

General notes on the agility and openness of the methods applied in this study can be found in the Appendix A. Now, it is time to have a closer look at the separate steps in the workflow:

1. Create a dataset.
   (a) Specify the text corpus to be used.
      i. Segment the corpus into smaller pieces, e.g. sentences.
      ii. Tokenize every segment.
   (b) Specify a gold standard for hyponymy.
   (c) Define hyponymy pairs for every segment in the corpus.
2. Split the dataset into subsets for training, validation and evaluation.
3. Implement a transformer language model (Vaswani et al. 2017) for hyponymy extraction.
   (a) Train the model on the training subset.
   (b) Use the validation subset to make sure that the model inferred general rules instead of memorizing input patterns.
4. Evaluate the model using the evaluation dataset.

Since machine learning is the major approach to be considered here, the size of the dataset (especially the training subset) is crucial (Hestness et al. 2017, p. 13). The rule of maximizing training data applies not only to machine learning, but also to other methods of distributional semantics, e.g. in the case of semantic relations (Herbelot and Vecchi 2015, p. 27). Therefore, it seems reasonable to prefer the Corpus Corporum (about 190 million tokens of plain text, cf. Roelli (2014)) over PROIEL (about 220.000 tokens of annotated text). Unfortunately,

this limits the usage of linguistic research results (i.e. annotations) as input for the language models. On the other side, the practical application of the model for language learning purposes becomes easier and more flexible, in the sense that, theoretically, end users can supply any plain Latin text of their choice, without the need to provide additional annotations.

Before we can start to detect hyponyms, we need to induce word senses, especially for polysemous words. To make sure that fine-grained distinctions between different usages of the same word are recognized by the model, it should be evaluated on a test dataset that has the following structure:

- For each segment in a pair of segments (e.g. sentences),
- look at the target word(s) and
- predict whether it was used in the same sense as in the other segment.

The second step can be formalized by providing character indices referring to the position of the target word in a string representation of the segment. This works even for cases where a target word form occurs multiple times in the segment and we want to induce the word sense for just a single specific instance. The third step, i.e. prediction of semantic similarity, can be binary in this very basic operationalization: The two target words are either used in the same sense or not. This dichotomy consciously abstracts from more complicated evidence, e.g. where two instances are used in a very similar, but not quite the same sense. In a subsequent effort, we may try to associate the current instance with other segments where the word usage is similar, thus clustering occurrences of the same word sense together.

Since there are no semantically annotated authentic text corpora for Latin, the next best thing to use are grammar books and dictionaries: They often distinguish between various word senses and offer exemplary contexts for each one (Georges 1913; Niederau 2012). Some of them cite entire text passages (Short et al. 1879; Kühner et al. 1914; Menge et al. 2009) to support their classifications. Such references can be seen as annotations: If a dictionary lists various text passages as containing a specific word in the same sense, this is roughly equivalent to semantically annotated datasets where sense identifiers from static sense inventories are assigned to tokens in a text corpus. Now, we could proceed by looking at the largest entries in a dictionary, assuming that they contain the largest number of different word senses. This, however, would amount to accepting the dictionary editor's intuition, which is not a usage-based approach. Instead, we will find the most polysemous words in a corpus inductively, by looking at their usage contexts. In this view, the most polysemous word will be that which is used in the most diverse contexts.

A naive formalization of diverse contexts is to count the number of types that occur in the same sentence as a target type in the corpus. This disregards the absolute frequencies of the types in a corpus and will lead to a Zipfian distribution (Zipf 1936, p. VI) of target types. Thus, function words like *et, in, ut, est, non, cum* 'and, in, as, is, not, with' will dominate the ranking just

because of their immense absolute frequency in the corpus. We can limit their impact by randomly down-sampling their occurrences using an arbitrary upper bound of 100. Since our goal is to produce exercises for language learning, the potential hyponyms should be rather frequent in a given corpus (Ellis et al. 2013, p. 27; Robillard et al. 2014, p. 2). It is therefore reasonable to establish a minimum threshold for word frequency. This criterion interacts with Latin being a highly inflectional language, thus producing many morphologically different forms for the same lexeme (see Table 2).

| | PROIEL | French Treebank | Hamburg Dependency Treebank |
|---|---|---|---|
| Types per Lemma | 3.64 | 1.21 | 2.09 |

Table 2: Average number of types per lemma for Latin, French and German.

Thus, we will perform this initial experiment on professionally lemmatized texts, which restricts us to the PROIEL treebank. Using lemmata instead of types has the advantage of increased frequency counts per item (esp. in small corpora) and allows us to easily group contexts by lexeme. The downside here is that the diversity of contexts cannot be distinguished anymore for various word forms of the same lexeme, but we can reintroduce that distinction in a later step. The comparison of two contexts for a given target word will be made using representations of sentences in a CVS. This mapping from textual to numerical entities (i.e. vectors) is not straightforward, because it is unclear how exactly we should model the difference between short and long sentences that share roughly the same content, but with different wording.

If we pad all shorter sentences to an arbitrary upper bound (e.g. the length of the longest sentence in the corpus) using placeholder values, the model will be able to calculate the cosine similarity[1] (C. Luo et al. 2017, p. 3) between two sentence vectors as input. However, naively using a fixed value as the placeholder in right-padded sequences (Sachan et al. 2018, p. 386) is equivalent to assuming that, compared to very long sentences, every short sentence tends to repeat the same word over and over again after a certain time. From a linguistic perspective, this formalization seems rather unintuitive because it does not take into account the effects of text length on lexis (Golcher et al. 2011, p. 31; Ochab et al. 2019, p. 141), which may lead to a clustering of input sentences roughly by (original) sequence length.

## 5.2 Sensitivity to Lexical and Syntactic Variation

Against this backdrop, normalization of text length is probably just as problematic as the disruptive influence of text length in the first place. Possible alternatives consist in manual ad hoc annotation or trying to circumvent the

[1] Throughout this thesis, the terms *cosine similarity* and *cosine distance* will be used interchangeably. The distance is usually a value between -1 and 1. The similarity is usually calculated by subtracting the distance from 1.

problem. We will proceed by testing a model of semantic diversity that includes lexical and syntactic cues while also ignoring text length. By training a transformer language model (Vaswani et al. 2017, p. 9) on the raw text of the PROIEL corpus, we build a CVS into which single tokens can be embedded. Our implementation uses Positional Encoding (Conneau et al. 2019, p. 4) to explicitly analyze every word's position in a given sequence, thereby including basic syntactic notions in the model. Besides, the aforementioned distributional approach makes the representations lexically predictable, i.e. components of collocations will be represented next to each other in the CVS. To corroborate these hypotheses, we evaluate our model on a few arbitrarily selected example sentences from PROIEL that are lexically identical, but differ in their word order (see Table 3). As a cross check, we have calculated the cosine similarity between entirely identical sentences and verified that it equals 1, which is the maximum value.

The outlier value for the sentences 13883 and 11010 indicates a strong change in the token's representation depending on its position in this context. Morphologically, there is an unambiguous agreement between the noun *possessiones* 'possessions' and its modifying adjective *multas* 'many' in case, number and gender. Thus, further research is to be done here to explain this outcome by testing for other possible interferences, e.g. sentence length, part of speech or lexeme-specific syntactic preferences. It seems that, at least for the PROIEL corpus, the relatively free word order does not automatically imply a semantic equivalence of various syntactic configurations (Devine et al. 2006, p. 452), even though some of the previous scholarly literature suggested this to be the case (Niemeyer et al. 1997, p. 421).

Lexis, on the other hand, is a much more obvious indicator of semantic differences, especially from a distributional perspective. To compare sentences with identical syntax, we arbitrarily pick examples from PROIEL that are longer than 3 tokens and differ in exactly 1 word, which was replaced by either an entirely different word or by a morphological variant of the same lexeme. Such syntactically almost identical sentences can be compared by computing the cosine similarity of their vector representations. These vectors are aggregated by averaging the vectors for all tokens in the sentence, which is a common strategy (Adi et al. 2016, p. 1). Again, as a cross check, we make sure that the cosine similarity for completely identical input sentences is almost at the maximum value. Then, we proceed to compare each sentence to its syntactical twin where 1 token was replaced (see Table 4).

A major problem with the underlying CVS is that its dimensions are not interpretable. In language modeling with neural networks, it is generally assumed that the vector representations entail all kinds of linguistic information, e.g. on morphology, syntax or semantics (Gladkova et al. 2016, p. 8; Rogers et al. 2017, p. 143). However, in each of the linguistic domains, fine-grained nuances may not have been captured adequately, such as the distinction of synonyms from other semantically related words (Karan et al. 2012, p. 115) or semantic analogies beyond single cases and categories (Rogers et al. 2017, 142f.). Therefore, it remains unclear why some substitutions invoke much greater changes in

similarity than others, e.g. the change of mood (sentences 20141 and 63139) vs. the change of speaker and perspective (sentences 18848 and 48253). An intuitive explanation relates to homographs, i.e. *veniam* could also be an accusative noun ('mercy') instead of a subjunctive verb ('I might come'). Further research in this direction would need to consider alternative sentence representations, inspection of the different attention head layers for the replaced tokens, and systematic trials for specific replacement strategies (inflectional variants, synonyms etc.). For our purposes, it is enough to have a very basic empirical indication that the language model is sensitive to subtle changes in syntax and lexis, which is an important prerequisite for our evaluation of polysemy.

Besides, we may note that most of the replaced tokens seem to exhibit some kind of semantic relation, as can be seen in LOD collections like WordNet:

- synonymy
  - *initium* 'beginning' and *principium* 'origin' both belong to the same synset (`https://latinwordnet.exeter.ac.uk/api/synsets/n/04515071/lemmas/`), defined as 'the first part or section of something'.
  - *adprehendere* 'to seize' and *prehendere* 'to lay hold of' both belong to the synset 'take hold of so as to seize or restrain or stop the motion of' (`https://latinwordnet.exeter.ac.uk/api/synsets/v/00986697/lemmas/`).
- co-hyponymy
  - *secundus* 'second' and *septimus* 'seventh' are most certainly co-hyponyms in contexts where entities are being enumerated. The appropriate Wordnet synset (`https://latinwordnet.exeter.ac.uk/api/synsets/a/02103770/lemmas/`) contains *septimus*, *primus* 'first' and *quintus* 'fifth' and is defined as 'being or denoting a numerical order in a series'. The fact that 'secundus' is not yet included there is probably due to the poor quality of the Latin WordNet (Franzini et al. 2019, p. 4). Note that there is no direct Latin equivalent for 'umpteenth' or other words denoting an arbitrary numerical sequence.
  - *pallidus* 'pale' and *niger* 'black' are most certainly co-hyponyms, as they belong to similarly defined synsets: '(used of color) having a dark hue' (`https://latinwordnet.exeter.ac.uk/api/synsets/a/00380983/`) and '(used of color) having a relatively small amount of coloring agent' (`https://latinwordnet.exeter.ac.uk/api/synsets/a/00380299/`).
  - *sicut* 'so as' and *tamquam* 'as much as' both belong to the same synset: '(often followed by äs) to the same degree' (`https://latinwordnet.exeter.ac.uk/api/synsets/r/00020994/lemmas/`).

We may conclude that lexical substitution without syntactic alteration often correlates with semantic relations between the two interchangeable lexemes. In contrast, the remaining two examples from Table 4 exhibit multiple inflections

of the same lemma (*venire* 'to come'), and a coreference pointing to the figure of Jesus (*me* 'me' and *Filium* 'Son') in the New Testament. In those cases, none of the traditional semantic relations is present. However, this approach in general is just an expert's intuition and has not been verified through arguments from actual language use. We therefore need to look for further evidence in the text corpus.

Finding empirical evidence for co-hyponymy is rather difficult in small text corpora, especially when adhering to the Distributional Inclusion Hypothesis mentioned above, which demands that the hyponyms and their hypernym share a large portion of their usage contexts (Geffet et al. 2005, p. 110). This ambitious requirement is often approximated using more readily available measures, such as graph density in distributional thesauri (Jana and Goyal 2018, p. 2; Jana, Varimalla, et al. 2020, p. 4).

Building on our heuristic approach from above, we may fine-tune the procedure by considering only the cases where at least one of the two replaced words is a noun, and by excluding all sentence pairs with mere morphological inflection instead of lexical substitution (see Table 5).

From these examples, we can see how the constraints lead to a higher precision, i.e. out of all results found, more and more tend to be instances of hyponymy or co-hyponymy. However, the recall (How many instances of hyponymy did we miss?) is unknown and can only be determined using gold-standard annotations. We will investigate that problem later in chapter XXX. Meanwhile, there are other problems to address. The first example in table 5 can be regarded at least as co-hyponymous: *feria* 'holiday' and *dies* 'day' both share a common transitive relation to the same WordNet synset ('a day reckoned from midnight to midnight', `https://latinwordnet.exeter.ac.uk/api/synsets/n/10878116/`). In our case, this interpretation is questionable and should be more strict: The hypernymy chain leads from *feria* (a holiday) to *dies* (any day from sunrise to sunset) and from there to the higher-level synset of 'a day reckoned from midnight to midnight'.

The second example is an instance of metonymy or, more specifically, *pars pro toto* (Gärdenfors 2015, p. 44): *nubes* 'cloud' and *caelum* 'sky' are both meant to refer to the same phenomenon, namely the sky above from where a voice is ringing out. In that sense, the two words are synonymous rather than co-hyponymous, let alone strictly hyponymous. There is no simple way for separating their relation from that in the first example, but neural networks may still be able to grasp various cues that, when considered in combination (Hinton et al. 2012, p. 1), enable us to clearly distinguish the two cases.

## 5.3 High Variance in Low-Frequency Types: A Case Study on Dis Legomena

In a first attempt to analyze possible confounding variables in our approach to Latin polysemy, we look at the occurrence frequencies for every lemma and

make two important observations (see Fig. 1):

- Lemmata with a higher frequency of occurrence tend to exhibit a higher average cosine distance between the vectors of their usage contexts.
- Lemmata with a lower frequency of occurrence have a higher variance, which leads to more outliers compared to the arithmetic mean. Conversely, highly frequent lemmata tend to be more predictable in their average cosine distance.

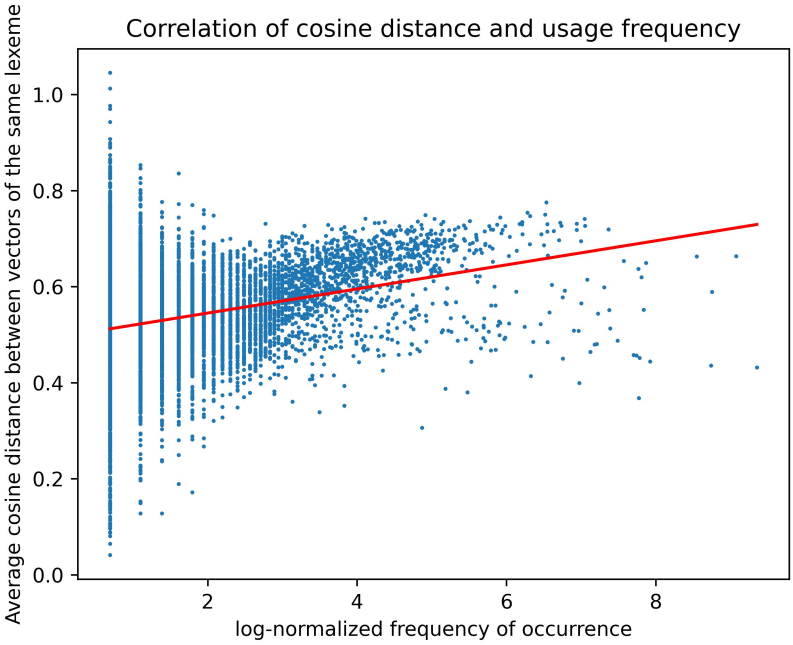


Figure 1: Correlation of average cosine distance and frequency of occurrence. Each blue dot represents the average cosine distance between all vectors for one lemma. Each vector was computed from one usage context of the lemma in the corpus. The red line indicates a linear regression of average cosine distance and usage frequency.

The second observation may serve as a reminder that we have to be particularly cautious when looking at low-frequency lemmata: Their relatively high variance in average cosine distance indicates a widely varying degree of polysemy. We can validate our operationalization of polysemy through cosine distance by manually inspecting conspicuous cases: Lemmata that occur twice in the whole corpus and only within the same sentence are, by definition, the least polysemous in our setting. The reasons for this behavior are as follows:

- Hapax Legomena have been discarded from the inspection because they only have one vector and cosine distance is not easily applicable to anything but pairs of vectors.
- The vectors are computed using distributional information (i.e. which other words appear in the sentence) and positional encoding. If a lemma appears twice in a sentence, the vectors for the two instances will differ mainly in their positional encoding, but not so much in the distributional information. This leads to very similar vectors with a low cosine distance.

Therefore, lemmata with 2 occurrences that both stem from the same sentence (e.g. *Augustofratensis, Constantinopolis* or *Sychem*, found in PROIEL 57662, 57780 and 23738, respectively) are prime examples for a low degree of polysemy. The phenomenon of low-frequency words tending to appear multiple times in quick succession (instead of a smooth distribution over the whole corpus) is referrd to as burstiness. It often results from a topical focus, which in turn influences lexis (Pierrehumbert 2012, p. 99). This is also true for PROIEL, as we can quickly determine from a simple calculation: 1,271 lemmata are Dis Legomena, i.e. they occur exactly twice in the corpus. 37 of them occur only within a single sentence. Thus, the observed probability of encountering both instances of a Dis Legomenon in a specific sentence is 37 / 19,335 = 0.002 , where 19,335 is the total number of sentences in the corpus. If, on the other hand, we take up a theoretical position and naively expect an equal probability of occurrence for Dis Legomena in each sentence, our expectation would look like this:

- In the whole corpus, there are 19,335 sentences with 217,797 tokens, which in turn derive from 8,872 lemmata.
- On average, each sentence contains about 11 tokens. Each of them could belong to any lemma. The probability of meeting an instance of a specific lemma at a specific place in the sentence is 1 / 8,872 = 0.0001 . An average sentence could be modeled as a sequence of Bernoulli trials, where the chance of success (i.e. meeting a specific lemma) at each token is 0.0001 . The chance of meeting that same lemma exactly twice in a sentence is then $P(k) = \binom{n}{k} \times p^k \times q^{n-k}$, where $n$ is the number of trials, $k$ is the desired number of successes, $p$ is the probability of success and $q$ the one of failure.
- Thus, $P(k) = \binom{11}{2} \times 0.0001^2 \times 0.9999^{11-2} = 55 \times 0.00000001 \times 0.9991 = 0.0000005$. Since we want to exclude any lemmata that are not Dis Legomena, we have to divide that result by a factor that corresponds to the fraction of Dis Legomena compared to all lemmata, i.e. 1,271 / 8,872 = 0.14 . As a consequence, our theoretical chance of encountering both instances of a Dis Legomenon in the same sentence becomes even smaller (0.00000007).

Now, we can compare that theoretical value (0.00000007) to the actually observed probability of encountering both instances of a Dis Legomenon within

the same sentence (0.002), which is almost 30,000 times higher. This immense deviation from our expectations leads to the conclusion that Bernoulli Trials are not a good model for lexis in sentences of this Latin corpus. We may even go further and argue that many statistical measures cannot be reasonably applied to language corpora because their mathematical preconditions are not met by the data (Shadrova 2020, p. 110). In particular, the words in a corpus are not distributed equally and their relative frequencies do not converge after a certain amount of time (Dębowski 2018, p. 5).

The first observation, on the other hand, suggests a positive correlation between polysemy and usage frequency: Words that we use more frequently tend to be more flexible in their meaning from a distributional point of view (S. Crossley et al. 2010, p. 576). We can verify that claim by subsampling a specific amount of usage contexts for every type in the corpus and calculating the TTR (Kettunen 2014, p. 229; Hamilton et al. 2016, p. 1) for them:

1. Select all types from PROIEL with an occurrence frequency $\geq$10.
2. For every selected type, collect all tokens that occurred in the same sentences as that type.
3. Omit all types did not occur together with at least 100 tokens.
4. For the remaining types, randomly sample 100 tokens from their usage contexts.
5. From the selected tokens, calculate the TTR for each type.

The thresholds of 10 and 100 were arbitrarily chosen. Their main purpose is to ensure that we have enough (i.e., 10) instances of a single type to avoid the large variance that became apparent in Fig. 1 and to mitigate the influence of text length effects (max. 100 tokens per type), which is a known problem for TTR and other measures of lexical diversity (Koizumi 2012, p. 66). Nevertheless, in the lower left corner of Fig. 2 (TTR $<$0.6, frequency $<$4), we can still observe some traces of higher lexical variance in low-frequency types. Generally, as indicated by the linear regression, high-frequency types tend to exhibit a slightly larger lexical diversity in their usage contexts. This underlines the importance of explicitly dealing with issues of polysemy in the extraction of hyponyms for language learning, as was outlined above.

## 5.4 Clustering of Contextual Embeddings

As we outlined before, the actual clustering of usage contexts for every token in the PROIEL corpus will be performed on the lexeme level. The procedure is as follows:

1. For every token in the corpus, compute its embedding vector in the sentence context.

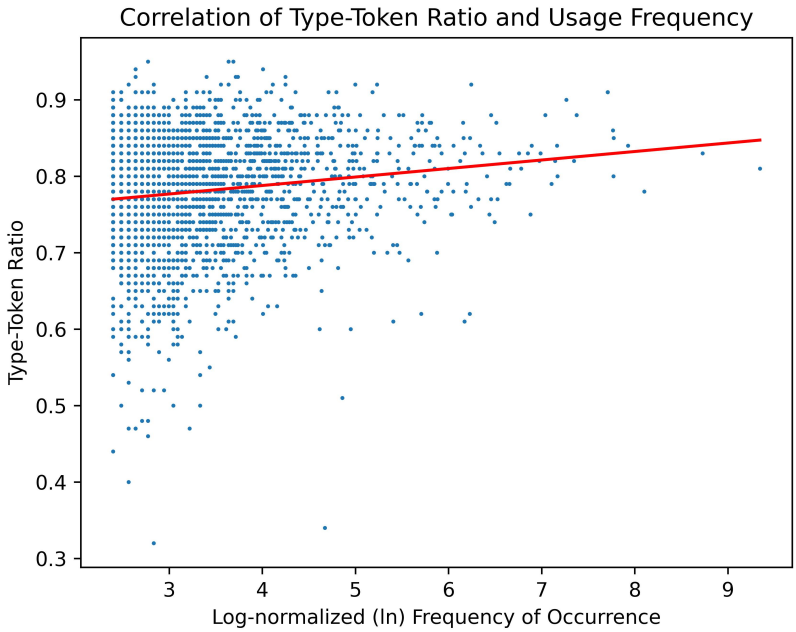


Figure 2: Correlation of TTR and frequency of occurrence. Each blue dot represents a specific type. TTRs were calculated only for types that occur at least 10 times in the corpus and have at least 100 tokens in their aggregated usage contexts, i.e., sentences. For the chosen types, a random sample of 100 tokens was selected for the calculation of TTR to avoid text length effects. The red line represents a linear regression.

2. Initially, treat every embedding vector as its own cluster.
3. For every pair of clusters, compute the similarity between their vectors. If a cluster contains more than one vector, compute the average for all of its vectors and use that for comparison to the other cluster.
4. Merge those clusters that are very close to each other.
5. Repeat steps 3 and 4 until a specific threshold is reached.

Ideally, the recursive iterations should take place after every merging of two clusters. This way, the newly formed cluster has a chance to be merged again immediately (if its similarity to another cluster is still above the threshold). However, for computational reasons, we instead use small batches of 15 clusters and compute the similarity values between all distinct pairs for them. This leads to about 15,000 batches being simultaneously processed in every iteration. Thus, the clustering method loses some of its accuracy while being substantially faster to compute.

The initial number of clusters, where each usage context represents its own cluster, is 217,923 , which corresponds closely to the number of tokens in the PROIEL corpus (217,797). By using a threshold of 0.9 for the cosine similarity that is required to merge two clusters, exactly 1 new cluster emerges. Correspondingly, the overwhelming majority of clusters was not merged at all. The merged cluster contains two usage contexts of the word *et* 'and':

(PROIEL 11203) *et* *responde-ns* *Iesus* *ai-t* *ill-is*
and reply-PART.PRES Jesus say.IMPF-3SG that-PL.DAT/ABL
and Jesus replied to them by saying

(PROIEL 16513) *et* *responde-ns* *Iesus* *ai-t* *ill-i*
and reply-PART.PRES Jesus say.IMPF-3SG that-DAT
and Jesus replied to him by saying

The two sentences differ only in their lexis, and only in the last token. Even for those last tokens, the relevant lemma (*ille* 'he, that') is the same. Thus, the representations are so close to each other (cf. Table 4) that they get merged into a larger cluster quite easily.

This changes when we set the similarity threshold to 0.8 . To achieve a higher performance of the clustering algorithm, we do not perform this additional processing step on the original clusters that were all unmerged. Instead, we use the results of the first clustering phase, which is not important in this early phase (because it barely differs from the original clusters), but will be crucial later on. This procedure yields 41 merged clusters, some of which contain up to 4 usage contexts. All in all, the merged vectors now include 88 (0.04%) usage contexts. Interestingly, the merges occurred not just for further instances of *et* 'and' (e.g., PROIEL 10565 and 48782) or other function words like *ibi* 'there' (PROIEL 57777 and 78440), but also for seemingly unrelated pairs like *videam – diligentiam* 'I may see – carefulness':

(PROIEL 77240)

*itaque exspect-o Thessalonic-ae act-a Kal Sext ex*
thus wait-1SG Thessaloniki-LOC transaction-PL first.day august from
*qu-ibus statu-a-m in tu-os ne agr-os*
which-PL.ABL decide-SBJV-1SG into your-PL.ACC.M whether field-ACC.PL
*confugi-a-m ut ne que vide-a-m homin-es qu-os*
flee-SBJV-1SG so.that not =and see-SBJV-1SG man-PL who-PL.ACC.M
*nol-i-m et te ut scrib-is vide-a-m et*
be.unwilling-SBJV-1SG and you.ACC as write-2SG see-SBJV-1SG and
*prop-ius si-m si quid ag-a-tur id*
near-COMP be.SBJV-1SG if something conduct-SBJV-3SG.PASS this.N
*que intelle-x-i cum t-ibi tum Quint-o*
=and perceive-PERF-1SG when you-DAT then Quintus-DAT/ABL
*fratr-i placere an abe-a-m Cyzic-um*
brother-DAT please or go.away-SBJV-1SG Cyzicus-ACC

Thus, I wait at Thessaloniki for the transactions of August 1. On the basis of them, I will decide whether to take refuge at your estate, so that I may not have to see those that I do not want to, and instead see you (as you wrote to me), and be closer if anything happens (and this, I have learned, is pleasant for you and especially for Quintus), or to go away to Cyzicus

(PROIEL 77314)

*in qu-o si iam nostr-a salus cum hac leg-e*
in which-ABL.N/M if already our-F salvation with this.F.ABL law-ABL
*despera-t-a eri-t vel-i-m*
give.up-PART.PERF.PASS-F.SG/N.PL be.FUT-3SG want-SBJV-1SG
*pro tu-o in me amor-e hanc*
for.the.sake.of your-DAT/ABL.M/N towards I.ACC love-ABL this.F.ACC
*inan-em me-am diligentia-m miserabil-em potius quam*
useless-ACC.M/F my-F.ACC diligence-ACC pitiful-ACC.M/F rather than
*inept-am put-e-s sin es-t aliquid spe-i*
silly-F.ACC think-SBJV-2SG if.however be-3SG anything hope-GEN/DAT
*d-e-s opera-m ut ma-ior-e diligenti-a posthac*
give-SBJV-2SG effort-ACC so.that large-COMP-ABL diligence henceforth
*a nostr-is magistrat-ibus defend-a-mur*
by our-PL.DAT/ABL official-PL.DAT/ABL defend-SBJV-1PL.PASS

Regarding this matter, if my salvation will already be hopeless together with this law, I would like you to think, because of your love towards me, that my useless diligence is pitiful rather than silly; if, however, there is any kind of hope, please try to make sure that, in the future, we will be defended by our officials with greater diligence

The passages belong to Cicero's Letter to Atticus, book 3, letters 15 and 23. They share almost no vocabulary except for function words like *tuus* 'your', *ut* 'that, so that', or *si* 'if'. Their semantic similarity is possibly related to the overlap of involved persons (Cicero and Atticus), but Cicero's brother Quintus is mentioned in the first passage. Another factor is the narrative structure:

both passages outline two potential future scenarios. In the first letter, Cicero ponders on his next action:

(a) go to Atticus' estate

(b) go to Cyzicus

Correspondingly, the second letter shows two potential outcomes that Cicero wishes for:

(a) Atticus should feel pity for him (if all hope is relinquished)

(b) Atticus should get the officials on Cicero's side (if he gets the chance)

The distribution of cluster size seems to follow a Zipfian distribution (see Fig. 3).

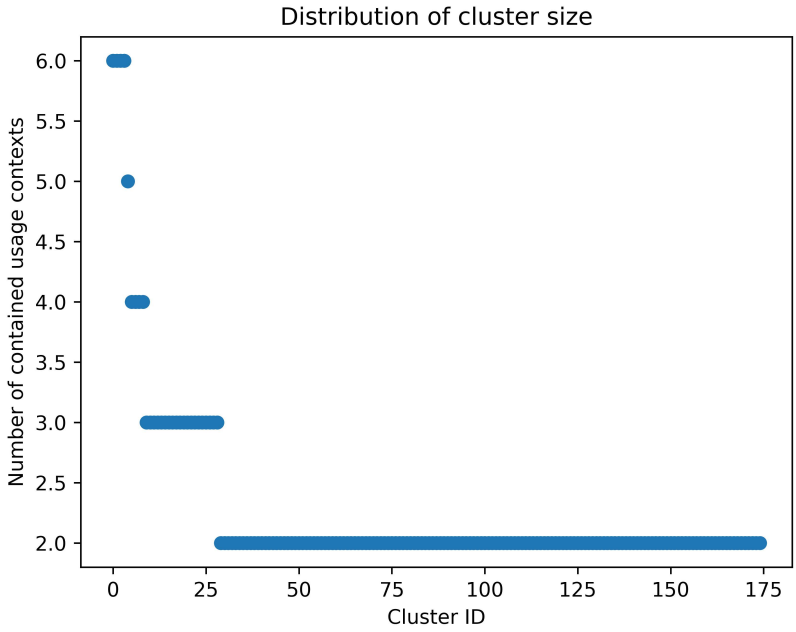


Figure 3: Distribution of cluster size for merged clusters at a similarity threshold of 0.9 . The largest cluster (128 usage contexts) was excluded for better legibility.

All merged clusters consist of usage contexts for the same lemma, i.e., there are no cross-lexeme clusters. This does not change if we lower the similarity threshold to 0.8,

| Context | | Cosine Similarity |
|---|---|---|
| (PROIEL 12966) | *si* ***fu-eri-t*** *oculus tu-us …*<br>if be.PF-SBJV-3SG eye your-M<br>if your eye was ... | |
| (PROIEL 17274) | *si oculus tu-us* ***fu-eri-t*** *…*<br>if eye your-M be.PF-SBJV-3SG<br>if your eye was ... | 0.81 |
| (PROIEL 13425) | ***qui*** *autem non habe-t …*<br>who but not have-3SG<br>but who does not have ... | |
| (PROIEL 17880) | *ab e-o autem* ***qui*** *non habe-t …*<br>from this-SG.ABL.M/N but who not have-3SG<br>but from him who does not have ... | 0.76 |
| (PROIEL 13883)<br>*era-t enim habe-ns* ***mult-as*** *possession-es*<br>be.IMPF-3SG for have-PART.PRES many-F.PL possession-PL.NOM/ACC<br>for he had many possessions | | |
| (PROIEL 11010)<br>*era-t enim habe-ns possession-es* ***mult-as***<br>be.IMPF-3SG for have-PART.PRES possession-PL.NOM/ACC many-F.PL<br>for he had many possessions | | 0.37 |
| (PROIEL 14258)<br>*… praeteri-bi-t* ***haec*** *generatio donec omn-ia haec*<br>... pass.by-FUT-3SG this.F.SG / N.PL generation until all-N.PL this.F.SG / N.PL<br>*fi-a-nt*<br>happen-SBJV-3PL<br>... this generation will pass by before all this happens | | |
| (PROIEL 18059)<br>*… praeteri-bi-t generatio* ***haec*** *donec omn-ia fi-a-nt*<br>... pass.by-FUT-3SG generation this.F.SG / N.PL until all-N.PL happen-SBJV-3PL<br>... this generation will pass by before everything happens | | 0.83 |
| (PROIEL 14376) | *opus* ***bon-um*** *opera-t-a es-t in me*<br>deed good-N do-part.pf-f be-3SG towards I.ACC<br>she has done a good deed towards me | |
| (PROIEL 11310) | ***bon-um*** *opus opera-t-a est in me*<br>good-N deed do-part.pf-f AUX towards I.ACC<br>she has done a good deed towards me | 0.75 |
| (PROIEL 11311) | *semper enim* ***pauper-es*** *habe-tis vo-bis cum*<br>always for poor-PL.F/M have-2PL you-abl with<br>for you always have poor people with you | |
| (PROIEL 19419) | ***pauper-es*** *enim semper habe-tis vo-bis cum*<br>poor-PL.F/M for always have-2PL you-abl with<br>for you always have poor people with you | 0.81 |

Table 3: Cosine similarity for tokens in almost identical sentences from PROIEL. Target tokens are highlighted in bold.

| Context | | Cosine Similarity |
|---|---|---|
| (PROIEL 34155) | *ego sum Alpha et Omega* ***initium*** *et finis*<br>I be.1SG Alpha and Omega beginning and end<br>I am Alpha and Omega, **beginning** and end | |
| (PROIEL 33474) | *ego sum Alpha et Omega* ***principium*** *et finis*<br>I be.1SG Alpha and Omega origin and end<br>I am Alpha and Omega, **origin** and end | 0.90 |
| (PROIEL 33726) | *et* ***secundus*** *angelus tub-a cecini-t*<br>and second angel trumpet-ABL sound.pf-3SG<br>and the **second** angel sounded his trumpet | |
| (PROIEL 33823) | *et* ***septimus*** *angelus tub-a cecini-t*<br>and seventh angel trumpet-ABL sound.pf-3SG<br>and the **seventh** angel sounded his trumpet | 0.88 |
| (PROIEL 33661) | *et ecce equus* ***pallidus***<br>and there horse pale<br>and, see, there is a **pale** horse | |
| (PROIEL 33655) | *et ecce equus* ***niger***<br>and there horse black<br>and, see, there is a **black** horse | 0.85 |
| (PROIEL 14139) | *dilig-e-s proxim-um tu-um* ***sicut*** *te ips-um*<br>love-FUT-2SG neighbor-ACC your-ACC like you.ACC self-ACC<br>you shall love your neighbor **like** yourself | |
| (PROIEL 11217) | *dilig-e-s proxim-um tu-um* ***tamquam*** *te ips-um*<br>love-FUT-2SG neighbor-ACC your-ACC as you.ACC self-ACC<br>you shall love your neighbor **as** yourself | 0.92 |
| (PROIEL 20141) | *si sic e-um vol-o manere donec* ***veni-am*** *quid ad te*<br>if so this-M.ACC want-1SG stay until come-SBJV.1SG what to you.ACC<br>if I want him to stay this way until I **come**, what is it to you | |
| (PROIEL 63139) | *si sic e-um vol-o manere donec* ***veni-o*** *quid ad te*<br>if so this-M.ACC want-1SG stay until come-1SG what to you.ACC<br>if I want him to stay this way until I **come**, what is it to you | 0.47 |
| (PROIEL 19306) | *quaere-ba-nt ergo e-um* ***prendere***<br>seek-IMPF-3PL therefore this-M.ACC catch<br>therefore, they wanted to **get a hold of** him | |
| (PROIEL 18951) | *quaere-ba-nt ergo e-um* ***adprehendere***<br>seek-IMPF-3PL therefore this-M.ACC seize<br>therefore, they wanted to **seize** him | 0.81 |
| (PROIEL 18848) | *qui cred-it in* ***me*** *habe-t vita-m aetern-am*<br>who believe-3SG in I-ACC have-3SG life-ACC eternal-F.ACC<br>who believes in **me** has an eternal life | |
| (PROIEL 48253) | *qui cred-it in* ***Fili-um*** *habe-t vita-m aetern-am*<br>who believe-3SG in Son-ACC have-3SG life-ACC eternal-F.ACC<br>who believes in **the Son** has an eternal life | 0.95 |

Table 4: Cosine similarity for sentences from PROIEL in which exactly one token was replaced. Replacements and their translations are highlighted in bold.

| Context | |
|---|---|
| (PROIEL 60041) | *sext-a* ***feri-a*** *in Syon*<br>sixth-F.SG.NOM holiday-F.SG.NOM in Syon<br>the sixth holiday [is celebrated] in Syon |
| (PROIEL 64252) | *sext-a* ***feri-a*** *in Syon*<br>sixth-F.SG.NOM holiday-F.SG.NOM in Syon<br>the sixth holiday [is celebrated] in Syon |
| (PROIEL 12797) | *et ecce vox de* ***cael-is*** *dice-ns*<br>and there voice from heaven-N.PL.ABL say-PART.PRES<br>and there a voice from the sky was saying |
| (PROIEL 13725) | *et ecce vox de* ***nub-e*** *dice-ns*<br>and there voice from cloud-ABL say-PART.PRES<br>and there a voice from the cloud was saying |
| (PROIEL 12929) | *nonne et* ***publican-i*** *h-oc faci-unt*<br>not also tax.gatherer-PL this-N.ACC do-3PL<br>do not tax gatherers also do this? |
| (PROIEL 12931) | *nonne et* ***ethnic-i*** *h-oc faci-unt*<br>not also pagan-PL this-N.ACC do-3PL<br>do not pagans also do this? |
| (PROIEL 13048) | *et a-it ill-i* ***Iesus***<br>and say-PF.3SG that-M.DAT Jesus<br>and Jesus told him |
| (PROIEL 11219) | *et a-it ill-i* ***scriba***<br>and say-PF.3SG that-M.DAT scribe<br>and the scribe told him |
| (PROIEL 13055) | *et responde-ns* ***centurio*** *a-it*<br>and answer-PART.PRES commander say-PF.3SG<br>and as a response, the commander said |
| (PROIEL 10471) | *et responde-ns* ***e-is*** *a-it*<br>and answer-PART.PRES that-M.PL.DAT say-PF.3SG<br>and responding to them, he said |
| (PROIEL 47503) | *vulp-es fovea-s habe-nt et volucr-es cael-i* ***tabernacul-a***<br>fox-PL pit-PL.ACC have-3PL and bird-PL sky-GEN tent-PL.ACC<br>the foxes have pits and the birds in the sky have tents |
| (PROIEL 62941) | *vulp-es fovea-s habe-nt et volucr-es cael-i* ***nid-os***<br>fox-PL pit-PL.ACC have-3PL and bird-PL sky-GEN nest-PL.ACC<br>the foxes have pits and the birds in the sky have nests |

Table 5: Nouns in almost identical sentences from PROIEL. Target tokens are highlighted in bold.

# Chapter 6

# Appendix

## 6.1 Appendix A: Agile and Open Methodology

This study will follow an approach which is borrowed from computer science, namely the agile methodology for software development (Fowler et al. 2001, p. 32). Coding will be used to establish a proof of concept for this study by implementing the discussed models and testing them on real-world corpus data. Furthermore, the basic notion of incremental and iterative progress also applies to the way the models themselves are being created: The dissertation will see a range of approaches and considerations being applied for various purposes, which reflects my own thoughts about the detection of semantic relations growing from initial, very basic foundations (e.g. plain contingency tables) to more complex environments of semantic study. The application of agile methodology to linguistic research is motivated not just by the related coding activities, but also by the similarities between dissertations and software products: In this view, a dissertation can be seen as a product to be delivered to customers, i.e. the research community. The analogy is enhanced further by the overlapping quality factor of openness: Dissertations are necessarily products of open science in much the same way as open source software is the product of other communities sharing the desire for unrestricted public access to creative commons (García-Penalvo et al. 2010, p. 518).

To demonstrate the viability of this transfer, the dissertation is split into various modules, separating, e.g., the bibliography from the rest of the documents and employing cross-references to put it all together for presentational purposes. These modules are, to some extent, directly visible in the sense of multiple source files that may be compiled to form a more complex document. Also, they are all publicly accessible at and subject to version control, thus enabling readers to trace the resulting thoughts back to their very roots. Eventually, the reason to make all these considerations explicit is not so much to suggest a high quality of the work process, but rather to emphasize their possibly tremendous influence on the results and how they are gathered.

# Bibliography


Adi, Yossi et al. (2016). "Fine-Grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks". In: *arXiv preprint arXiv:1608.04207*. arXiv: 1608.04207.

Akbik, Alan, Duncan Blythe, and Roland Vollgraf (2018). "Contextual String Embeddings for Sequence Labeling". In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649.

Amrami, Asaf and Yoav Goldberg (May 2019). "Towards Better Substitution-Based Word Sense Induction". In: *arXiv:1905.12598 [cs]*. arXiv: 1905.12598 [cs].

Anstatt, Tanja (2009). "Typen Semantischer Relationen". In: *Die Slavischen Sprachen. Ein Internationales Handbuch Zu Ihrer Geschichte, Ihrer Struktur Und Ihrer Erforschung*. Ed. by T. Berger et al., pp. 906–915.

Ayşe, Şerbetçi, Orhan Zeynep, and Pehlivan İlknur (June 2011). "Extraction of Semantic Word Relations in Turkish from Dictionary Definitions". In: *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 11–18.

Baïdak, Nathalie, Marie-Pascale Balcon, and Akvile Motiejunaite (2017). "Key Data on Teaching Languages at School in Europe. 2017 Edition. Eurydice Report." In: *Education, Audiovisual and Culture Executive Agency, European Commission*. ISSN: 9294924823.

Baker, Collin F. and Christiane Fellbaum (Aug. 2009). "WordNet and FrameNet as Complementary Resources for Annotation". In: *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*. Suntec, Singapore: Association for Computational Linguistics, pp. 125–129.

Bartunov, Sergey et al. (2016). "Breaking Sticks and Ambiguities with Adaptive Skip-Gram". In: *Artificial Intelligence and Statistics*, pp. 130–138.

Bengio, Yoshua et al. (2003). "A Neural Probabilistic Language Model". In: *Journal of machine learning research* 3.Feb, pp. 1137–1155.

Boleda, Gemma, Marco Baroni, and Louise McNally (2013). "Intensionality Was Only Alleged: On Adjective-Noun Composition in Distributional Semantics". In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013): Long Papers; 2013 Mar 20-22; Postdam, Germany.*

*Stroudsburg (USA): Association for Computational Linguistics (ACL)*. ACL (Association for Computational Linguistics), pp. 35–46.

Bone, Daniel et al. (2015). "Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises". In: *Journal of autism and developmental disorders* 45.5, pp. 1121–1136. ISSN: 0162-3257.

Borges Völker, Emanuel et al. (Aug. 2019). "HDT-UD: A Very Large Universal Dependencies Treebank for German". In: *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*. Paris, France: Association for Computational Linguistics, pp. 46–57. DOI: 10.18653/v1/W19-8006.

Carlson, Andrew et al. (2010). "Toward an Architecture for Never-Ending Language Learning". In: *Twenty-Fourth AAAI Conference on Artificial Intelligence*, pp. 1306–1313.

Coenen, Hans Georg (Feb. 2013). *Analogie und Metapher: Grundlegung einer Theorie der bildlichen Rede*. Walter de Gruyter. ISBN: 978-3-11-089463-9.

Conneau, Alexis and Guillaume Lample (2019). "Cross-Lingual Language Model Pretraining". In: *Advances in Neural Information Processing Systems* 32, pp. 7059–7069.

Crossley, Scott, Tom Salsbury, and Danielle McNamara (2010). "The Development of Polysemy and Frequency Use in English Second Language Speakers". In: *Language Learning* 60.3, pp. 573–605. ISSN: 1467-9922. DOI: 10.1111/j.1467-9922.2010.00568.x.

Crossley, Scott A., Tom Salsbury, and Danielle S. McNamara (2010). "The Development of Semantic Relations in Second Language Speakers: A Case for Latent Semantic Analysis". In: *Vigo International Journal of Applied Linguistics* 7, pp. 55–74.

Dai, Andrew M., Christopher Olah, and Quoc V. Le (July 2015). "Document Embedding with Paragraph Vectors". In: *arXiv:1507.07998 [cs]*. arXiv: 1507.07998 [cs].

Dębowski, Łukasz (2018). "Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited". In: *Entropy* 20.2, p. 85. DOI: 10.3390/e20020085.

Devine, Andrew M. and Laurence D. Stephens (2006). *Latin Word Order: Structured Meaning and Information*. Oxford University Press. ISBN: 0-19-972050-9.

Devlin, Jacob et al. (June 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Divjak, Dagmar and Stefan Th Gries (2009). "Corpus-Based Cognitive Semantics: A Contrastive Study of Phasal Verbs in English and Russian". In: *Studies in cognitive corpus linguistics*, pp. 273–296.

Ellis, Nick C., Matthew Brook O'Donnell, and Ute Römer (2013). "Usage-Based Language: Investigating the Latent Structures That Underpin Acquisition". In: *Language Learning* 63, pp. 25–51.

Faruqui, Manaal et al. (2016). "Problems with Evaluation of Word Embeddings Using Word Similarity Tasks". In: *arXiv preprint arXiv:1605.02276*. arXiv: 1605.02276.

Fellbaum, Christiane and Piek Vossen (June 2012). "Challenges for a Multilingual Wordnet". In: *Language Resources and Evaluation* 46.2, pp. 313–326. ISSN: 1574-0218. DOI: 10.1007/s10579-012-9186-z.

Firth, J. R. (1957). "A Synopsis of Linguistic Theory 1930-55". In: *Studies in Linguistic Analysis*. Ed. by J. R. Firth. Oxford: Blackwell, pp. 1–32.

Fischer, Andreas et al. (2010). "Ground Truth Creation for Handwriting Recognition in Historical Documents". In: *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pp. 3–10. DOI: https://dl.acm.org/doi/pdf/10.1145/1815330.1815331.

Fowler, Martin and Jim Highsmith (2001). "The Agile Manifesto". In: *Software Development* 9.8, pp. 28–35.

Franzini, Greta et al. (2019). "Nunc Est Aestimandum: Towards an Evaluation of the Latin WordNet". In: *CLiC-it*.

Garcia-Gamez, Ana B and Pedro Macizo (2019). "Learning Nouns and Verbs in a Foreign Language: The Role of Gestures". In: *Applied Psycholinguistics* 40.2, pp. 473–507. ISSN: 0142-7164.

García-Penalvo, Francisco J., Carlos García de Figuerola, and Jose A. Merlo (2010). "Open Knowledge Management in Higher Education". In: *Online Information Review* 34.4, pp. 517–519.

Gärdenfors, Peter (2015). "Semantic Transformations". In: *Language and Semiotic Studies* 1.1, pp. 41–51.

Gardner, D. (Apr. 2007). "Validating the Construct of Word in Applied Corpus-based Vocabulary Research: A Critical Survey". In: *Applied Linguistics* 28.2, pp. 241–265. ISSN: 0142-6001, 1477-450X. DOI: 10.1093/applin/amm010.

Geffet, Maayan and Ido Dagan (2005). "The Distributional Inclusion Hypotheses and Lexical Entailment". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 107–114.

Georges, Karl Ernst (1913). *Ausführliches Lateinisch-Deutsches Handwörterbuch*.

Gladkova, Anna, Aleksandr Drozd, and Satoshi Matsuoka (2016). "Analogy-Based Detection of Morphological and Semantic Relations with Word Embeddings: What Works and What Doesn't". In: *Proceedings of the NAACL Student Research Workshop*, pp. 8–15.

Golcher, Felix and Marc Reznicek (2011). "Stylometry and the Interplay of Topic and L1 in the Different Annotation Layers in the FALKO Corpus". In: *Proceedings of Quantitative Investigations in Theoretical Linguistics 4*. Ed. by Amir Zeldes and Anke Lüdeling. Humboldt-Universität zu Berlin, pp. 29–34.

Grefenstette, Edward and Mehrnoosh Sadrzadeh (2011). "Experimental Support for a Categorical Compositional Distributional Model of Meaning". In: *arXiv preprint arXiv:1106.4058*. arXiv: 1106.4058.

Gries, Stefan Th and Dagmar Divjak (2009). "Behavioral Profiles: A Corpus-Based Approach to Cognitive Semantic Analysis". In: *New directions in cognitive linguistics*, pp. 57–75.

Gyllensten, Amaru Cuba, Ariel Ekgren, and Magnus Sahlgren (2019). "R-Grams: Unsupervised Learning of Semantic Units in Natural Language". In: *Proceedings of the 13th International Conference on Computational Semantics - Student Papers*, pp. 52–62.

Hagiwara, Masato, Yasuhiro Ogawa, and Katsuhiko Toyama (2009). "Supervised Synonym Acquisition Using Distributional Features and Syntactic Patterns". In: *Information and Media Technologies* 4.2, pp. 558–582.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky (2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". In: *arXiv preprint arXiv:1605.09096*. arXiv: 1605.09096.

Harris, Zellig S. (Aug. 1954). "Distributional Structure". In: *Word* 10.2-3, pp. 146–162. ISSN: 0043-7956, 2373-5112. DOI: 10.1080/00437956.1954.11659520.

Haug, Dag TT and Marius Johndal (2008). "Creating a Parallel Treebank of the Old Indo-European Bible Translations". In: *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pp. 27–34.

Herbelot, Aurélie and Mohan Ganesalingam (2013). "Measuring Semantic Content in Distributional Vectors". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2, pp. 440–445.

Herbelot, Aurélie and Eva Maria Vecchi (2015). "Building a Shared World: Mapping Distributional to Model-Theoretic Semantic Spaces". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 22–32.

Hestness, Joel et al. (Dec. 2017). *Deep Learning Scaling Is Predictable, Empirically*.

Hinton, Geoffrey E et al. (2012). "Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors". In: *arXiv preprint arXiv:1207.0580*. arXiv: 1207.0580.

Huang, Eric et al. (July 2012). "Improving Word Representations via Global Context and Multiple Word Prototypes". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, pp. 873–882.

Jana, Abhik and Pawan Goyal (2018). "Network Features Based Co-Hyponymy Detection". In: *arXiv preprint arXiv:1802.04609*. arXiv: 1802.04609.

Jana, Abhik, Nikhil Reddy Varimalla, and Pawan Goyal (2020). "Using Distributional Thesaurus Embedding for Co-hyponymy Detection". In: *arXiv preprint arXiv:2002.11506*. arXiv: 2002.11506.

Jimeno-Yepes, Antonio J., Bridget T. McInnes, and Alan R. Aronson (June 2011). "Exploiting MeSH Indexing in MEDLINE to Generate a Data Set for Word Sense Disambiguation". In: *BMC Bioinformatics* 12.223, pp. 1–14. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-223.

Karakanta, Alina, Jon Dehdari, and Josef van Genabith (2018). "Neural Machine Translation for Low-Resource Languages without Parallel Corpora". In: *Machine Translation* 32, pp. 167–189.

Karan, Mladen, Jan Šnajder, and Bojana Dalbelo Bašic (2012). "Distributional Semantics Approach to Detecting Synonyms in Croatian Language". In: *Information Society*, pp. 111–116.

Kettunen, Kimmo (2014). "Can Type-Token Ratio Be Used to Show Morphological Complexity of Languages?" In: *Journal of Quantitative Linguistics* 21.3, pp. 223–245. ISSN: 0929-6174. DOI: https://www.tandfonline.com/doi/pdf/10.1080/09296174.2014.911506.

Koizumi, Rie (2012). "Relationships between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens". In: *Vocabulary Learning and Instruction* 1.1, pp. 60–69.

Kühner, Raphael and Carl Stegmann (1914). *Ausführliche Grammatik Der Lateinischen Sprache, 2. Teil: Satzlehre*. Vol. 1. Hanovre.

Lee, Jin Ha and Xiao Hu (2012). "Generating Ground Truth for Music Mood Classification Using Mechanical Turk". In: *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 129–138. DOI: https://dl.acm.org/doi/pdf/10.1145/2232817.2232842.

Lewis, Martha (Sept. 2019). "Compositional Hyponymy with Positive Operators". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria: INCOMA Ltd., pp. 638–647. DOI: 10.26615/978-954-452-056-4_075.

Liu, Yang et al. (2011). "Combining Integrated Sampling with SVM Ensembles for Learning from Imbalanced Datasets". In: *Information Processing & Management* 47.4, pp. 617–631. ISSN: 0306-4573.

Luo, Chunjie et al. (Oct. 2017). "Cosine Normalization: Using Cosine Similarity Instead of Dot Product in Neural Networks". In: *arXiv:1702.05870 [cs, stat]*. arXiv: 1702.05870 [cs, stat].

Luo, Fuli et al. (July 2018). "Incorporating Glosses into Neural Word Sense Disambiguation". In: *arXiv:1805.08028 [cs]*. arXiv: 1805.08028 [cs].

McGillivray, Barbara (2013). *Methods in Latin Computational Linguistics*. Leiden; Boston: Brill.

Menge, Hermann, Thorsten Burkard, and Markus Schauer (2009). *Lehrbuch Der Lateinischen Syntax Und Semantik*. Fourth. Wissenschaftliche Buchgesellschaft. ISBN: 3-534-13661-6.

Mihalcea, Rada (June 2008). *SemCor Corpus*.

Mihalcea, Rada, Timothy Chklovski, and Adam Kilgarriff (July 2004). "The Senseval-3 English Lexical Sample Task". In: *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain: Association for Computational Linguistics, pp. 25–28.

Mikolov, Tomas, Kai Chen, et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv preprint arXiv:1301.3781*, pp. 1–12. arXiv: 1301.3781.

Mikolov, Tomas, Ilya Sutskever, et al. (2013). "Distributed Representations of Words and Phrases and Their Compositionality". In: *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Minozzi, S. (2010). "The Latin WordNet Project". In: *Latin Linguistics Today: Akten Des 15. Internationalen Kolloquiums Zur Lateinischen Linguistik, Innsbruck, 4.- 9. April 2009*. Ed. by Peter Anreiter and M. Kienpointner. Vol. 137. Innsbrucker Beiträge Zur Sprachwissenschaft. Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck, Bereich Sprachwissenschaft, pp. 707–716. ISBN: 978-3-85124-723-7.

Moro, Andrea et al. (2014). "Annotating the MASC Corpus with BabelNet". In: *LREC*, pp. 4214–4219.

Navigli, Roberto and Simone Paolo Ponzetto (July 2010). "BabelNet: Building a Very Large Multilingual Semantic Network". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 216–225.

Niederau, Philipp (2012). *Navigium Latein-Deutsch-Wörterbuch*.

Niemeyer, Jürgen and Herwig Krenn (1997). "Zur Stellung Des Attributiven Adjektivs Im Lateinischen Und in Den Romanischen Sprachen. Syntaktische Gemeinsamkeiten Und Unterschiede". In: *Semiotische Prozesse Und Natürliche Sprache: Festschrift Für Udo L. Figge Zum 60. Geburtstag*. Ed. by Andreas Gather and Heinz Werner. Stuttgart: Franz Steiner Verlag, pp. 411–423.

Nivre, Joakim et al. (2016). "Universal Dependencies v1: A Multilingual Treebank Collection". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1659–1666.

Ochab, Jeremi K. and Holger Essler (2019). "Stylometry of Literary Papyri". In: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, pp. 139–142.

Ono, Masataka, Makoto Miwa, and Yutaka Sasaki (2015). "Word Embedding-Based Antonym Detection Using Thesauri and Distributional Information". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 984–989.

Peters, Matthew E. et al. (2018). "Deep Contextualized Word Representations". In: *arXiv preprint arXiv:1802.05365*. arXiv: 1802.05365.

Pierrehumbert, Janet B (2012). "Burstiness of Verbs and Derived Nouns". In: *Shall We Play the Festschrift Game? Essays on the Occasion of Lauri Carlson's 60th Birthday*. Ed. by Diana Santos, Krister Linden, and Wanjiju Ng'ang'a. Springer, pp. 99–115.

Pinkal, Manfred (1993). "Semantik". In: *Einführung in Die Künstliche Intelligenz*. Ed. by Günther Görz. First. Addison-Wesley, pp. 425–498.

Pons Bordería, Salvador (2014). "Paths of Grammaticalization in Spanish o Sea". In: *Discourse and Pragmatic Markers from Latin to the Romance Languages*. Ed. by Chiara Ghezzi and Piera Molinelli, pp. 109–136.

Punuru, Janardhana and Jianhua Chen (Feb. 2012). "Learning Non-Taxonomical Semantic Relations from Domain Texts". In: *Journal of Intelligent Information Systems* 38.1, pp. 191–207. ISSN: 1573-7675. DOI: 10.1007/s10844-011-0149-4.

Robillard, Manon et al. (2014). "Monolingual and Bilingual Children with and without Primary Language Impairment: Core Vocabulary Comparison". In: *Augmentative and alternative communication* 30.3, pp. 267–278.

Roelli, Philipp (2014). "The Corpus Corporum, a New Open Latin Text Repository and Tool". In: *Archivum Latinitatis Medii Aevi-Bulletin du Cange (ALMA)*.

Rogers, Anna, Aleksandr Drozd, and Bofang Li (2017). "The (Too Many) Problems of Analogical Reasoning with Word Vectors". In: *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pp. 135–148.

Roller, Stephen, Katrin Erk, and Gemma Boleda (2014). "Inclusive yet Selective: Supervised Distributional Hypernymy Detection". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1025–1036.

Sachan, Devendra Singh et al. (2018). "Effective Use of Bidirectional Language Modeling for Transfer Learning in Biomedical Named Entity Recognition". In: *Machine Learning for Healthcare Conference*, pp. 383–402.

Schibel, Wolfgang (2013). "Zur Aneignung Lateinischer Literatur Und Sprache". In: *Forum Classicum*, pp. 113–124.

Seddah, Djamé et al. (2018). "Cheating a Parser to Death: Data-driven Cross-Treebank Annotation Transfer". In: *Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Shadrova, Anna Valer'evna (July 2020). "Measuring Coselectional Constraint in Learner Corpora: A Graph-Based Approach". In: DOI: 10.18452/21606.

Short, Charles and Charlton T. Lewis (1879). *A Latin Dictionary*. Oxford: Clarendon Press.

Souillé-Rigaut, Chris (2010). "A Semantic Account of Quasi-Lexemes in Modern English-Processing Semiotic Units of Greek or Latin Origin into Lexical Units". PhD thesis. University of Kansas.

Stoll, Sabine et al. (2012). "Nouns and Verbs in Chintang: Children's Usage and Surrounding Adult Speech". In: *Journal of Child Language* 39.2, pp. 284–321. ISSN: 1469-7602.

Taslim, Fadilla (2014). "An Experimental Study of Teaching Vocabulary by Using Hyponymy Games on the Seventh Grader F MTs Syech Ibrahim Payakumbuh". In: *Al-Ta lim Journal* 21.3, pp. 189–197.

Tengi, Randee I. (1998). "Design and Implementation of the WordNet Lexical Database and Searching Software". In: *WordNet: an electronic lexical database*. Ed. by Christiane Fellbaum, pp. 105–127.

Uzuner, Özlem et al. (Sept. 2010). "Community Annotation Experiment for Ground Truth Generation for the I2b2 Medication Challenge". In: *Journal*

*of the American Medical Informatics Association* 17.5, pp. 519–523. ISSN: 1067-5027. DOI: 10.1136/jamia.2010.004200.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 5998–6008.

Wei, Qiong and Roland L. Dunbrack Jr (July 2013). "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics". In: *PLOS ONE* 8.7, e67863. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0067863.

York, James and Jonathan William deHaan (2018). "A Constructivist Approach to Game-Based Language Learning: Student Perceptions in a Beginner-Level EFL Context". In: *International Journal of Game-Based Learning (IJGBL)* 8.1, pp. 19–40.

Yuan, Dayu et al. (Nov. 2016). "Semi-Supervised Word Sense Disambiguation with Neural Models". In: *arXiv:1603.07012 [cs]*. arXiv: 1603.07012 [cs].

Yuan, Sylvia, Cynthia Fisher, and Jesse Snedeker (2012). "Counting the Nouns: Simple Structural Cues to Verb Meaning". In: *Child Development* 83.4, pp. 1382–1399. ISSN: 1467-8624. DOI: 10.1111/j.1467-8624.2012.01783.x.

Zipf, George Kingsley (1936). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*.