ASRAEL - Acquisition of Semantic RelAtions bEtween Latin nouns

Konstantin Schulz
Humboldt-Universität zu Berlin

XXXabstractXXX

Contents

List of Figures

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. Specifically, I hope that an improved integration of hyponymy into the learning process can lead to a higher level of language skills:

> In cognitive linguistics, the word itself with its network of polysemous senses came to be regarded as a category in which the senses of the word (i.e. the members of the category) are related to each other by means of general cognitive principles such as metaphor, metonymy, generalization, specialization, and image-schema transformations. (Nerlich and Clarke 2003, p. 5)

If there is a parallel between semantic relations and the way that humans think, it seems reasonable to exploit that connection for learning purposes. Unfortunately, this still leaves us with the lack of a clearly defined research object:

> Another problem arising from polysemy and homonymy is lexical ambiguity, and the precise relationship between polysemy, homonymy, ambiguity and vagueness is still an unresolved issue in lexical semantics. (Nerlich and Clarke 2003, p. 4)

It is therefore important to define the various semantic relations and distinguish between them as clearly as possible. In the case of, e.g., synonymy, a good operationalization is needed to retrieve the relevant instances from a corpus of text data:

> Polysemy requires the researcher to determine whether two usage events are identical or sufficiently similar to be considered a single sense, what the degree of similarity is between different senses, where to connect a sense to others in the network, and which sense(s) to recognize as prototypical one(s). [...] in addition, [linguists] have to decide what the differences are between the near-synonyms as well as what the relation is between semantically similar words in a domain. (Divjak and Gries 2009, p. 274)

In my study, I will follow the approach of distributional semantics, describing specific semantic relations between words in terms of their co-occurrences and typical contexts in a corpus. Previous research has been carried out using, e.g., association measures to detect synonymy:

> The value of distributional features $f_j^D(x, z)$ is determined so that it represents the degree of commonality of context c_j shared by the word pair (x, z). [...] The advantage of this feature construction is that, given the independence assumption between word x and z , the feature value is easily calculated as the simple sum of two corresponding pointwise mutual information weights as: $f_j^D(x, z) = PMI(x, c_j) + PMI(z, c_j)$ [...]. (Hagiwara, Ogawa, and Toyama 2009, p. 566)

One of the critical parts in that view is the term "word pair", which introduces the linguistically ill-defined concept

of word into an otherwise seemingly precise mathematical formula. Thus, the usage of bare token-based PMI (Pointwise Mutual Information) for distributional questions may lead to severely skewed results:

> [...] strong collocation effects can influence the measurement of information negatively: it is an open question which phrases should be considered 'words-with-spaces' when building distributions. (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. Besides, polysemy itself 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:

> Some of the errors we observe may also be related to word senses. For instance, the word medium, to be found in the pair magazine – medium, can be synonymous with middle, clairvoyant or again mode of communication. In the sense of clairvoyant, it is clearly more specific than in the sense intended in the test pair. As distributions do not distinguish between senses, this will have an effect on our results. (Herbelot and Ganesalingam 2013, p. 444)

Since the solution to this problem may be to disambiguate word senses for every single token in a corpus, 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.

Hyponymy

After dealing with various issues of polysemy, some of the second-tier relations may have to be redefined when looked at from a distributional perspective, e.g. hyponymy: "Although beverage is an umbrella word for many various types of drinks, speakers of English use it in very particular contexts. So, distributionally, it is not a 'general word'." (Herbelot and Ganesalingam 2013, p. 443) In this view, other existing taxonomies that may have acted as models of hypernymy need to be replaced by a definition that reflects how actual usage contexts can limit our selection of possible candidates for hypernymy:

> [The Distributional Inclusion Hypothesis] states that more specific terms appear in a subset of the distributional contexts in which more general terms appear. So, animal can occur in all the contexts in which dog can occur, plus some contexts in which dog cannot – for instance, rights can be a typical cooccurrence for animal (e.g. "animal rights"), but not so much for dog (e.g. #"dog rights"). (Roller, Erk, and Boleda 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:

> It is worth noting that the skipgram model uses subsampling of common words, which is an optimization introduced to compensate for the power law distribution in common vocabularies. Also, the skipgram model controls for collocations by dampening the impact of frequent collocations. (Gyllensten, Ekgren, and Sahlgren 2019, p. 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 creating word embeddings using the skipgram model, 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:

> Theoretischer Anwendungsbereich eines Wortes: die Klasse der Gegenstände, auf die ein Wort kraft seiner Bedeutung beschreibend angewandt werden kann (TA). Empirischer Anwendungsbereich eines Wortes: die Menge der Gegenstände, zu deren Beschreibung ein Wort - bis zu einem gewissen Zeitpunkt - in der Erfahrung eines bestimmten SprachteiInehmers bereits angewandt wurde (EA). (Coenen 2013, p. 46)

This may be transferred to the domains of corpus linguistics and language learning: Given a text corpus that includes a learner's desired knowledge (e.g. all the literature one wants to read), the theoretical scope of a certain word would be equivalent to all the distinct relevant contexts in the corpus in which that word occurs. In contrast, the empirical scope of that same word would be equal to the number of distinct relevant contexts in the corpus in which it occurs and which are known to the learner (at any given point in time of his learning process). Thus, the theoretical scope will remain static as long as we do not change the corpus, while the empirical one will always be dynamic in any case. From this point of view, we can also restrict our notion of semantic relations: Hyponymy, for instance, will be tied to those word networks that are present in a specific corpus and, for a learner, specific instances of hyponomy relations will not exist unless they have already been part of a learning situation. 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 build lexical knowledge. In order to accomplish that first part, we need to have clearly defined requirements to qualify a relation as one of hyponymy:

> Auf der konzeptuellen Ebene, auf der auch Similarität und Kontiguität angesiedelt sind, beruht Hyponymie auf der taxonomischen Inklusion von Konzepten. Die taxonomische Inklusion ist nichts anderes als eine Ober-/Unterbegriffsbeziehung [...]. Diese Konzeptpaare sind teilidentisch, weil die Extension (die Menge der mit dem Begriff gemeinten Phänomene) des Unterbegriffs in der des Oberbegriffs enthalten ist und gleichzeitig die Intension (die Menge der dem Begriff zugeordneten Eigenschaften) des Oberbegriffs in der des Unterbegriffs enthalten ist [...]. (Gévaudan 1999, p. 16)

This notion of semantic super- and subordination between words closely resembles the principles in cognitive linguistics cited above Nerlich and Clarke 2003, p. 5: If hyponymy

Konzeptuelle Relationen	Semantische Techniken der Rede
Identität	(Synonymie)
Taxonomische Inklusion	Spezifizierung / Generalisierung
Taxonomische Exklusion	kohyponymische Übertragung
Similarität	Metapher (Vergleich)
Kontiguität	Metonymie (Bezugsetzung)

Abb. 1: "Techniken der Rede" und konzeptuelle semantische Relationen

Figure 1. Hyponymy as generalisation/specification (Gévaudan 1999, p. 18)

refers to some underlying taxonomy in which a hyponym is subordinate to its hypernym, our most important task will be to establish that taxonomy in a corpus-based manner. Therefore, lexical resources like the Latin WordNet (Minozzi 2010) may contribute a reference model for such a purpose because, apart from its own invention of synsets, it includes a built-in notion of hyponymy. Unfortunately, the WordNet defines such relations based on intuition, not on actual language use. It is therefore static and non-empirical. However, sind it was supposedly constructed by experts, we can use their assessment for evaluating the corpus-based extraction of hyponymy. This kind of approach is supported by recent studies:

> The WE-T model receives supervised information from synonym and antonym pairs in thesauri and infers the relations of the other word pairs in the thesauri from the supervised information. The WE-TD model incorporates corpus-based contextual information (distributional information) into the WE-T model, which enables the calculation of the similarities among in-vocabulary and out-of-vocabulary words. [...] Our WE-TD model achieved the highest score among the models that use both thesauri and distributional information. (Ono, Miwa, and Sasaki 2015, p. 984-988)

According to Ono, Miwa, and Sasaki 2015, it is not sufficient to use thesauri like the WordNet for detecting semantic relations; rather, we should use that as a basis and provide additional information from actual corpora. In both cases, we can increase the amount of provided information by relying on associated properties of hyponymy relations, e.g. transitivity:

> Die Relationen der Hyper- und Hyponymie sind transitiv. Wenn A Hyponym in Bezug auf B ist, dann auch in Bezug auf die Hyperonyme von B, und wenn B Hyperonym in Bezug auf A ist, dann auch in Bezug auf die Hyponyme von A. (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* (engl. 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 use morphological hints to get faster access to information about hyponymy: "[...] zur Bildung von Hyponymen kann z.B. das Suffix -ovye/-evye (seld' ‚Hering' – sel'devye ‚Heringsartige') verwendet werden (s. Ginzburg 1985, 9)." (Anstatt 2009, p. 913) This is especially true for Latin with its relatively rich morphology: the meaning of *artifex* (engl. artist) may be associated directly to *ars* (engl. art) and *facere* (engl. to make), without necessarily traversing a hierarchical chain of increasing generalisation. 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:

> The head constituent in both compounds is on the right, the modifier constituent on the left, and, although 'cardiopathy' is not paraphrasable, we can still deduce from its semiotic units that the compound as a whole is a hyponym of the meaning carried by the semiotic unit -path-. (Souillé-Rigaut 2010, p. 33)

This works similarly for Greek loanwords in Latin, e.g. *xylophytum* (a certain kind of plant) cannot be 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. Besides, such a model is able to explain why we can immediately see semantic structure in newly coined words:

> 'Dontopedalogy is the science of opening your mouth and putting your foot in it, a science which I have practiced for a good many years'. There is evidence that we are dealing with a secondary compound of the type W + X + Y, that is to say, with the semiotic units -dont- + -ped- + -log- being concatenated synchronically. [...] secondary compounds W + XY are always expanded primary compounds in which WXY is a hyponym of XY. (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.

Corpus data

General notes on the agility and openness of the methods that will be applied in this study can be found in the Appendix A. Now, it is time to have a closer look at the involved resources, most notably the corpus.

Appendix

Appendix A: Agile and Open Methodology

This study will follow an approach which is borrowed from computer science, i.e. the agile methodology for software development:

> [...] agile processes assume and encourage the alteration of requirements while the code is being written. As such, design cannot be a purely up-front activity to be completed before construction. Instead, design is a continuous activity that's performed throughout the project. (Fowler and Highsmith 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. pure 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.

> We think that Open Knowledge comprises Open Software, Open Content, Open Science and

> Open Innovation. [...] Open Software owes its deepest roots to Open Access; Open Contents are related to open access to the educative, cultural or divulgative contents that are published under a non restrictive license that allows copy and distribution, but also the right to modify works. Open Science is devoted to the open access to scientific contents, while Open Innovation transfers the Open Access principles to the enterprise production world, which is actually indispensable for the enhancement of University-Enterprises relationships. (García-Penalvo, García de Figuerola, and Merlo 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.

References

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.

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

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.

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

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.

Gévaudan, Paul (1999). "Semantische Relationen in Nominalen Und Adjektivischen Kompositionen Und Syntagmen". In: *PhiN. Philologie im Netz* 9, pp. 11–34.

Gyllensten, Amaru Cuba, Ariel Ekgren, and Magnus Sahlgren (2019). "R-Grams: Unsupervised Learning of Semantic Units in Natural Language". en-us. 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.

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.

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.

Nerlich, Brigitte and David D. Clarke (2003). "Polysemy and Flexibility: Introduction and Overview". In: *Polysemy. Flexible Patterns of Meaning in Mind and Language*. Ed. by Brigitte Nerlich et al. Berlin: Mouton de Gruyter, pp. 3–30.

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.

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.

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.