Published in: Fuß, Eric/Konopka, Marek/Trawiński, Beata/Waßner, Ulrich H. (eds.): Grammar and corpora 2016. - Heidelberg: Heidelberg University Publishing, 2018. Pp. 425-434

Christian Lang, Roman Schneider, Karolina Suchowolec

Extracting Specialized Terminology from Linguistic Corpora

Abstract In this paper, we present our approach to automatically extracting German terminology in the domain of grammar using texts from the online information system grammis as our corpus. We analyze existing repositories of German grammatical terminology and develop Part-of-speech patterns for our extraction thereby showing the importance of unigrams in this domain. We contrast the results of the automatic extraction with a manually extracted standard. By comparing the performance of well-known statistical measures, we show how measures based on corpus comparison outperform alternative methods.

Keywords Grammatical terminology, terminological structures, automatic term extraction, grammatical information system

1 Introduction

The information system *grammis* (Schneider and Schwinn 2014) is an online resource on German grammar, hosted by the Institute for the German Language (IDS) in Mannheim. It comprises a wide range of specialist texts on grammatical phenomena of the German language. Additionally, *grammis* offers terminological resources: a dictionary for short reference, and a thesaurus organizing explicit relationships between terminological concepts for the automatic expansion of full-text queries. Established more than a decade ago, the whole system is currently being evaluated and re-designed. As for the current content, we observe that a broad spectrum of grammatical terminology used in the specialist hypertexts is covered neither by the dictionary nor by the thesaurus. We believe and will demonstrate that this coverage can be enhanced by applying automatic term extraction (ATE), i.e. the automatized identification and extraction of terms from domain-specific corpora.

We follow Heylen and De Hertog (2015) by adopting their characterization of a *term* as being part of the "core vocabulary of a specialised domain" (c.f. also Nakagawa and Mori 2002, Kaguera and Umino 1996 among others) which corresponds to German industry standards as defined by DIN2342. However, the classification of a specific entity as *term* (vs. *non-term*) is not a trivial task. Nazar (2016) points out that "in the absence of an intensional definition for the entity *term* researchers must resort to an operational definition" (Nazar 2016: 145), e.g. to a consultation of experts in the domain. In ATE, "the term/non-term categorisation [is] not binary but rather presented as a continuum, in the form of a list of candidates ranked according to a score that represents an estimate of the probability of the candidate being a term" (Nazar 2016: 145). Kageura and Umino (1996: 279f.) point out that the statistical methods used to identify and score term candidates share common assumptions based on the candidates' usage; one of those assumptions appearing more frequently in a specific domain than in general.[1] The quality of an ATE's statistical ranking of candidates can, then, be assessed by the degree to which it coincides with the manual evaluation of the expert.

There has been a substantial amount of research into ATE and its application, however mostly in technological domains (e.g. Nazar 2016, Lossio Ventura et al. 2014, Wermter and Hahn 2005, Frantzi et al. 2000). Zhang et al. (2008) compare different statistical measures applied in automatic term extraction tasks. Their comparative study in the domains of biology and medicine indicates that the domain has an "impact on the performance of ATR[2] algorithms" (Zhang et al. 2008: 2111). They also note that "[…] evaluation in other kinds of domains, notably less technical ones, have been lacking" (Zhang et al. 2008: 2109).

In this paper, we present our approach to extract relevant terminology in the domain of German grammar. As there is – to our knowledge – no evaluation study for this domain, we focus on a comparison of different algorithms. Hence, we implement an array of well-established statistical measures used in automatic term extraction tasks with an emphasis on contrasting corpus comparing measures with alternative measures. We evaluate the performance of the extraction algorithms by comparing the ATE's results to a standard manually extracted by a terminology/linguistics expert (MTE).

1 Kageura and Umino (1996: 280) also point out that while those assumptions seem reasonable, „the task of proper theorization is yet to be carried out."

2 Zhang et al. (2008) use the term Automatic Term Recognition (ATR) instead of Automatic Term Extraction (ATE).

2 Corpus

Our test set of *grammis* texts constitutes a corpus of 2,491 documents with a total of 1.2 million tokens and 44,000 types. Contents range from concise descriptions to more detailed discussions. From a technical point of view, all primary data and meta-data is coded within semi-structured XML instances that are composed of semantic markup elements ("title", "subtitle", "literature" etc.). As common in linguistic texts, most of the documents contain natural language example sentences for illustration purposes. These sentences, mostly taken from newspaper articles, are not consistently identified by semantic markup. This results in a substantial number of non-domain specific words which ATE has to handle.

3 Method

We start with standard linguistic preprocessing – applying TreeTagger (Schmid 1995), we assign Part-of-speech tags (POS) and stem the words in the corpora. After that, we apply three filters in order to block undesired candidates from extraction: the first filter exploits the semantic markup of the XML instances. In particular, it excludes bibliographical references and example sentences if they are marked as such. The second – statistical – filter is based on a comparison of our target corpus with a general domain reference corpus (see 3.2). A term candidate is eligible for extraction only if its relative frequency is higher in the specialized target corpus than in a general domain reference corpus (see Gelbukh et al. 2010). The statistical filter is implemented to minimize the amount of noise that is introduced by the non-terminological example sentences. No absolute frequency threshold is applied.[3] The third filter is based on POS patterns as described in 3.1. All candidates that satisfy the POS filter, the relative frequency threshold, and the semantic markup-filter are extracted from our target corpus.[4] They are subsequently ranked by the algorithms described in 3.2.

3 The manually extracted standard (see 4) includes a total of 67 hapax legomena with a frequency of 1, e.g. *Pseudocleft-Satz* ('pseudo cleft sentence').

4 Coordinated composites are a special challenge for extraction. Coordinated nouns share a morpheme that is omitted in one of them, e.g.: *Ereignis- und Betrachtzeit* ('event time and focus time'). Both, *Ereigniszeit* and *Betrachtzeit* are key terms, whereas the coordination is not. We extract the coordination and treat both coordinated elements as unigrams.

3.1 Linguistic Filter – POS Patterns

Justeson and Katz (1995) propose POS patterns for terminology extraction in English by analyzing dictionaries of different technical domains. The benefit of applying POS filters is the improvement of precision. The drawback is a potentially reduced recall. In order to minimize the risk of a reduced recall based on too narrow POS filters, we analyze the prevalent POS patterns of German grammatical terms in the above-mentioned *grammis* thesaurus and in the online version of the alphabetic index of *Duden – die Grammatik* (Duden 2017). The analysis of a total of 2,984 terms shows that 82 % of them are either nominal or adjectival unigrams, while only 15 % are bigrams of an adjective and a noun. These results contrast with Justeson and Katz (1995) who find that "the majority of technical terms do consist of more than one word" (Justeson and Katz 1995: 9); this observation, however, is based on English dictionaries in technical domains.

Our POS filter incorporates the following patterns that represent 99 % of the terms analyzed: N, A, AN, NN, N Prep N, N Det N, (V), A A N.[5]

3.2 Ranking Candidates

In order to rank the extracted candidates, we compare a series of well-established statistical measures that have been used in similar automatic term extraction tasks (see Heylen and De Hertog 2015 or Zhang et al. 2008 for an overview).[6] The implemented measures fall into one of two categories: measures based on corpus comparison and measures not based on corpus comparison. For the first type, our target corpus is compared to a randomly extracted sample from DeReKo (German Reference Corpus; Kupietz and Keibel 2009). It covers various text types and genres, and contains approx. 970,000 tokens and 80,000 types. In this group

5 **N**: nouns, proper names, numbers; **A**: adjectives, attributive and predicative; **Prep**: prepositions, **Det**: determiners, **V**: verbs. However, we exclude verbs from the extraction. With a share of a mere 0.34 % of the analyzed grammatical terms and a share of 11 % of the words in our target corpus, the inclusion of verbs would have increased noise for a minor improvement of recall.

6 Some of the measures we implemented are also used in the extraction of keywords (c.f. Heylen and De Hertog 2015: 219, also Kageura and Umino (1996) for a discussion of the close relation between the two fields). The measures we implemented have been used to extract terms in other (technological) domains, for example: **LL** by Gelbukh et al. (2010) for computer science, **Weird** by Gillam et al. (2007) for nanotechnology, **C-value** by Frantzi et al (2000) for medicine, **P-Mod** Wermter and Hahn (2005) for biomedicine. **TFIDF**, while prototypically applied in keyword extraction, is used by Zhang et al. (2008) as a baseline in their comparative study.

we implement the following measures: *Log-Likelihood based distance* – **LL** (Dunning 1993), *Simple Math* (with an *add-N parameter* of 10) – **SM_10** (Kilgarriff 2009) and *Weirdness* – **Weird** (Ahmad et al. 1999). All these measures evaluate a candidate's termhood (in the sense of Kageura and Umino 1996) and are based on the presumption that "terms are by definition domain-specific, and as a consequence are hypothesised to occur more frequently in their proper domain than they do in other domains or in general language use" (Heylen and De Hertog 2015: 219). While comparing the corpora, bigrams and trigrams are treated the same way as unigrams. Since bigrams and trigrams are generally less frequent, they are ranked lower in comparison to unigrams. For terms spanning more than one word, this is a crucial point in the analysis. The C-value and P-Mod measures (see below) are one way of incorporating information about the frequency of multi-word units and their relationship to the frequencies of shorter multi-word units contained in them.

The second type of measures is not based on corpus comparison. We implement three measures of this type: first, **TFIDF** (Spärck Jones 1972), which is widely used in text mining. TFIDF weighs a candidate's frequency in the corpus with its document frequency. Second, Frantzi et al.'s **C-value** (2000), which is based on frequency, and takes into consideration a candidate's likelihood of being nested in a construction. We use a modified version to account for unigrams (Lossio-Ventura et al. 2014). In the third place, we implement Wertmer and Hahn's paradigmatic modifiability, **P-Mod** (Wermter and Hahn 2005), also in a modified version to account for unigrams. Both C-value and P-Mod are hybrid approaches that combine a candidate's unithood and termhood (in the sense of Kageura and Umino 1996) and were both originally designed to identify multigram terms. In a final step, we also implement **t-value** to assess the unithood of multigrams and a distance metric based on longest common subsequence to detect spelling variants among the candidates. For calculating bonuses, we use semantic markups from the original XML files. Candidates receive a bonus of 30 % or 10 % if they are mentioned in a title or a subtitle respectively.

4 Results and Discussion

To evaluate our ATE results, we ask a linguistic terminologist to perform a manual terminology extraction from a randomly chosen subset of 120 out of the 2,491 documents in the corpus. The expert is asked to extract all linguistic terms regardless of structure, i.e., without POS-filtering. The results of this manual extraction serve as a gold standard for the quality of our ATE. We choose this design over a manual evaluation of the term candidates identified by the ATE as we want to prevent a bias towards parameters inherent to the ATE.

The manual extraction results in a list of 1,001 terms.[7] A large majority of 98 % are nouns, adjectives and their combination. 82.6 % of the manually extracted terms are unigrams, which corresponds to our analysis of existing repositories of German grammar described in 3.1. 948 of these standard terms are also found by ATE. With a total of 5,314 ATE candidates, this means a recall of 94.7 % with an overall precision[8] of 17.8 %.

The imperfect recall score cannot be attributed to the narrow POS-filter. Six terms in the standard are not in the scope of the ATE's POS filter; five of them are verbs. We observe that nominal and/or adjectival equivalents of all those verbs are retrieved by ATE. The main reason (27 %) for the imperfect recall score is a higher relative frequency in the general domain reference.

Regarding precision, the analysis of the top-ranked candidates missing in the standard shows at least five obvious key terms such as *flexion*, *outside field*, *phonological*, *unmarked* and *unstressed*. Besides, a candidate's spelling variants are sometimes treated differently by the expert: e.g., *Aufforderungsmodus* ('prompt mode') was deemed a term, whereas *Aufforderungs-Modus* was not. We attribute this to performance errors by the expert rather than to lack of expertise. In any case, this is a strong argument for always combining manual and automatic term extraction: the major advantage of manual extraction is the specialized knowledge of the expert; the brute force of ATE and its being based on objective corpus evidence can compensate for possible performance errors by the expert.

We further evaluate the precision of the ATE by ranking the candidates according to the implemented measures described in 3.2. Table 1 shows the ATE's precision for all implemented measures at various cutoffs, thus for the top i ranked candidates each. The results indicate that the precision of corpus-comparing measures is generally higher than the measures based on the target corpus only; *Weirdness* demonstrates the highest precision.

7 We retrospectively excluded a total of 28 terms from the standard. This was done either because of typos or because their exact form was not found in the documents. This applies primarily to complex NPs such as *local and temporal adverbials*. The expert extracted both *local adverbials* and *temporal adverbials*, even though the exact string *local adverbials* is not present in the text.

8 Recall is the fraction of terms that were successfully extracted: $R = \frac{\text{correctly extracted terms}}{\text{all standard terms}}$.

Precision is the fraction of extracted candidates that are terms: $P = \frac{\text{correctly extracted terms}}{\text{all extracted candidates}}$.

Table 1: Precision of ATE.

	Top i Ranked Candidates Evaluated			
Ranking Method	i = 50	i = 100	i = 500	i = 1000
Freq	56.0 %	60.0 %	45.2 %	38.5 %
TFIDF	76.0 %	68.0 %	51.8 %	42.9 %
Weird	96.0 %	88.0 %	67.6 %	51.4 %
SM_10	90.0 %	77.0 %	57.0 %	44.6 %
LL	78.0 %	75.0 %	57.6 %	45.4 %
C-value	66.0 %	68.0 %	51.0 %	39.9 %
P-Mod	58.0 %	62.0 %	46.6 %	38.3 %

Taking recall into account, Figure 1 displays precision-recall curves for all implemented measures. Increasing recall, the decrease in precision is slower for *Weirdness'* than for the other measures.

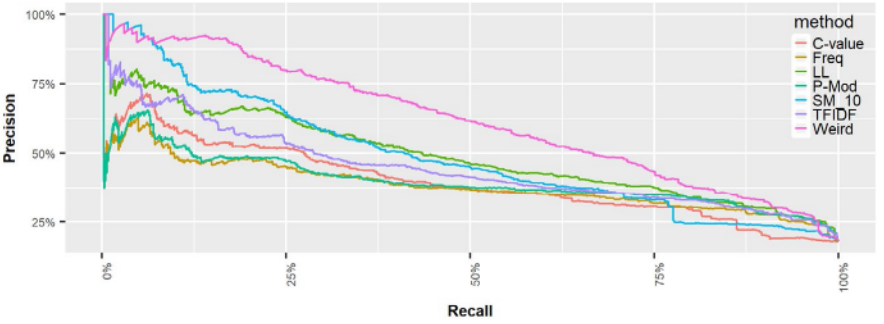

Figure 1: Precision/Recall graph.

As another metric to evaluate the ranking measures, we calculate the *Average Precision* (*AvP*) (Su et al. 2015) which is defined as:

$$\sum_{i=1}^{N} P(i)\Delta R(i)$$

In this formula, N represents the total number of candidates, P(i) is the precision at a cutoff of i candidates and ΔR(i) is the change in recall between cutoff i-1 and i. The AvP score is higher the more actual terms are among the higher ranked candidates. Figure 2 shows the AvP values for the examined measures:

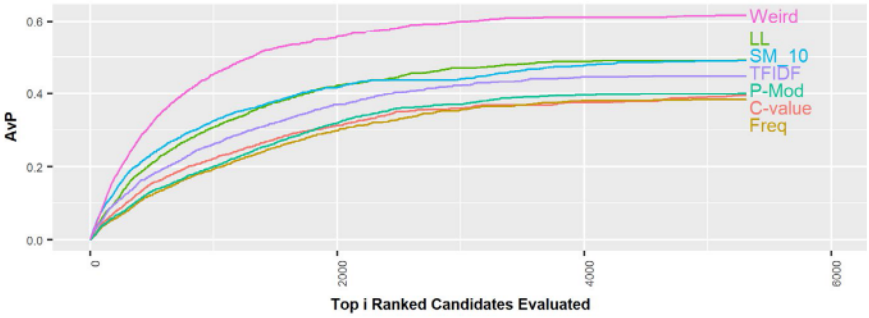

Figure 2: Average Precision (AvP).

Overall, the *Weirdness* measure shows the best performance; compared to the other measures, higher ranked candidates are more likely to be terms.[9] Measures that are based on corpus comparison outperform those that are based on the target corpus only. We attribute this result to the subset of high frequency candidates which are part of the general domain, e.g. *difference*, *example*. The comparison with a general language corpus results in a lower ranking of those candidates.[10] Finally, due to the high proportion of unigrams among the terms manually extracted by the expert, the algorithms that were designed to identify multigram terms show a weaker performance.

5 Concluding Remarks

We presented our approach to extract German grammatical terminology from linguistic corpora, and compared the performance of different ATE methods in this domain. The results indicate that corpus-comparing methods perform better than measures that are not based on corpus comparison. We showed the importance of unigrams in the domain of German grammar by analyzing both existing terminology repositories and the results of the manual extraction by an expert. This result contrasts with the prevalence of multigram terms in technical domains as stated by Nakagawa and Mori (2002) or Justeson and Katz (1999). The tendency towards shorter terms can be interpreted as characteristic for the domain of grammar, confirming Frantzi et al.'s (2000) observation that terms

9 In Zhang et al. (2008) *Weirdness* outperformed *TFIDF* and *C-value* when applied to the Wikipedia Corpus, however performed worse when applied to the life science corpus Genia.

10 Six of the ten most frequent candidates are words of the general domain. *C-value* and *P-Mod* rank five of them in their top ten.

tend to be shorter in arts compared to science and technology. Furthermore, German word formation allows for complex compound-unigrams that correspond to multiword units in English.

References

Ahmad, Khurshid, Lee Gillam and Lena Tostevin. 1999. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation, Retrieval (WILDER). In Ellen Voorhees and Donna Harman (eds.), *NIST Special Publication 500-246: The Eighth Text Retrieval Conference (TREC-8)*, 717—724. Gaithersburg, MA.

DIN 2342. 2011. *Begriffe der Terminologielehre*. Berlin: Beuth Verlag.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Journal of Computational Linguistics — Special Issue on Using Large Corpora*, 19(1): 61—74.

Duden. 2017. Grammatische Fachausdrücke. http://www.duden.de/sprachwissen/sprachratgeber/Grammatische-Fachausdrucke (February 02, 2018).

Frantzi, Katerina, Sophia Ananiadou and Hideki Mima. 2000. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. *International Journal on Digital Libraries* 3(2): 115—130.

Gelbukh, Alexander, Grigori Sidorov, Eduardo Lavin-Villa and Liliana Chanona-Hernandez. 2010. Automatic Term Extraction Using Log-Likelihood Based Comparison with General Reference Corpus. In *Proceedings of the 15th International Conference on Application of Natural Language Processing to Information Systems, NLDB 2010*, 248—255. Berlin/Heidelberg: Springer.

Gillam, Lee, Mariam Tariq and Kurshid Ahmad. 2007. Terminology and the Construction of Ontology. In Fidelia Ibekwe-SanJuan, Anne Condamines and M. Teresa Cabré Castellví (eds.), *Application-Driven Terminology Engineering*. Benjamins Current Topics 2: 49—73.

Heylen, Kris and Dirk De Hertog. 2015. Automatic Term Extraction. In Hendrik J. Kockaert and Frieda Steurs (eds.), *Handbook of Terminology. Volume 1*, 203—221. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Justeson, John S. and Slava M. Katz. 1995. Technical Terminology: some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering* 1(1): 9—27.

Kageura, Kyo and Bin Umino. 1996. Methods of Automatic Term Recognition: A Review. *Terminology* 3(2): 259—289.

Kilgarriff, Adam. 2009. Simple Maths for Keywords. In Michaela Mahlberg, Victorina González-Díaz and Catherine Smith (eds.), *Proceedings of Corpus Linguistics Conference CL2009*, University of Liverpool, UK, July 2009.

Kupietz, Marc and Holger Keibel. 2009. The Mannheim German Reference Corpus (DEREKO) as a Basis for Empirical Linguistic Research. In Makoto Minegishi and Yuji Kawaguchi (eds.), *Working Papers in Corpus-based Linguistics and Language Education*, No. 3., 53—59. Tokyo: Tokyo University of Foreign Studies (TUFS).

Lossio Ventura, Juan Antonio, Clement Jonquet, Mathieu Roche and Maguelonne Teisseire. 2014. Biomedical Terminology Extraction: A new Combination of Statistical and Web Mining Approaches. *Proceedings of Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2014)*, Paris, France.

Nakagawa, Hiroshi and Tatsunori Mori. 2002. A Simple but Powerful Automatic Term Extraction Method. *Proceedings of the Second International Workshop on Computational Terminology*, Stroudsburg, PA, USA: ACL: 1—7.

Nazar, Rogelio. 2016. Distributional Analysis Applied to Terminology Extraction. First Results from the Domain of Psychiatry in Spanish. *Terminology* 22(2): 141—170.

Schmid, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland: 1—9.

Schneider, Roman and Horst Schwinn. 2014. Hypertext, Wissensnetz und Datenbank: Die Webinformationssysteme grammis und ProGr@mm. In Franz Josef Berens and Melanie Steinle (eds.). *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 337—346. Mannheim: IDS.

Spärck Jones, Karen. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 28(1): 11—21.

Su Wanhua, Yan Yuan and Mu Zhu. 2015. <http://dx.doi.org/10.1145/2808194.2809481>. *Proceedings of the ACM SIGIR 2015 International Conference on the Theory of Information Retrieval*: 349—352.

Wermter, Joachim and Udo Hahn. 2005. Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*: 843—850.

Zhang, Ziqi, José Iria, Christopher Brewster and Fabio Ciravegna. 2008. A Comparative Evaluation of Term Recognition Algorithms. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*: 2108—2113. Marrakech: ELRA.