A Laypeople Study on Terminology Identification across Domains and Task Definitions

Anna Hätty
Robert Bosch GmbH
IMS, University of Stuttgart
anna.haetty@de.bosch.com

Sabine Schulte im Walde
Institute for Natural Language Proessing (IMS),
University of Stuttgart
schulte@ims.uni-stuttgart.de

Abstract

This paper introduces a new dataset of term annotation. Given that even experts vary significantly in their understanding of termhood, we offer a novel perspective to explore the common, natural understanding of what constitutes a term: Laypeople annotate single-word and multi-word terms, across four domains and across four task definitions. Analyses based on inter-annotator agreement offer insights into differences in term specificity, term granularity and subtermhood.

1 Introduction

Terms are linguistic units which characterize a specific topic domain, and their identification is relevant for a number of NLP tasks, such as information retrieval and automatic translation. Not only the automatic extraction of terms is a challenging task, but also their manual definition and identification: while we find a range of gold standard corpora for the evaluation of term extraction systems for English (Kim et al., 2003; Bernier-Colborne and Drouin, 2014; Zadeh and Schumann, 2016) and to a lesser extent also for German (Arcan et al., 2014; Arcan, 2017; Hätty et al., 2017), these benchmark datasets vary hugely in terms of granularity of term definition, topic and thematic focus. All datasets have in common that they have been annotated by domain experts and/or by terminologists, which is considered a necessary requirement for term evaluation (Castellví, 1999; Gouws et al., 2007). However, Estopà (2001) shows that even experts with different perspectives on terminology (e.g., terminologists, domain experts, translators and documentalists) vary significantly in their annotation of terms. Moreover, although individual studies describe different layers of terminology (Trimble, 1985; Roelcke, 1999), there is a lack of empirical studies. This raises the question whether there is a common, natural understanding of what constitutes a term, and to what extent this term is associated to a domain.

In this study, we examine the concept of terminology from a new perspective. Differently to previous annotation studies, we investigate judgments of laypeople, rather than experts, and specify on analyzing their (dis-)agreements on common assumptions and core issues in term identification: the word classes of terms, the identification of ambiguous terms, and the relations between complex terms and possibly included subterms. To ensure a broad understanding of term identification, we designed four different tasks to address the granularities of term concepts, and we performed all annotations across four different domains in German: diy, cooking, hunting, chess. Finally, we compare the annotations to the output of an unsupervised hybrid term extraction system.

2 Material and Tasks

Domains The data for term identification comprise German open-source texts from the websites *wikiHow*[1], *Wikibooks*[2] and *Wikipedia*. All texts have been pos-tagged with the *Tree Tagger* (Schmid, 1994); compound splitting was performed with *Compost* (Cap, 2014) and manually post-edited. In total, the text basis consists of 20 texts (five per domain) with ≈5 sentences each. All texts together contain 3,075 words, distributed over the following four domains:

- diy: *"do it yourself"* (708 words)
- cooking (624 words)
- hunting (900 words)
- chess (843 words)

[1] https://de.wikihow.com/
[2] https://www.wikibooks.org/

Proceedings of NAACL-HLT 2018, pages 321–326
New Orleans, Louisiana, June 1 - 6, 2018.

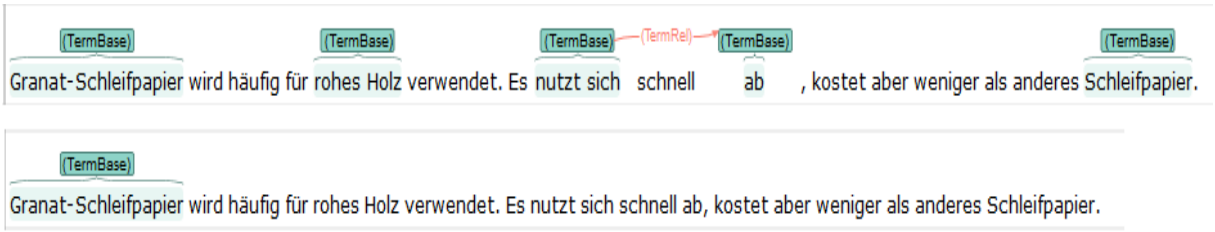

Figure 1: Example of WebAnno annotation for DS (top) and GL (bottom).

Term Identification Tasks In order to investigate the effect of term definition on their identification, we specified the following four tasks:

- highlight domain-specific phrases (**DS**)
- create an index (**IND**)
- define unknown words for creating a translation lexicon (**TR**)
- create a glossary (**GL**)

We assumed the four tasks to provide different strengths of associating the terms with the domains: DS and IND were expected to demand a broad range of terms that characterize the domains. TR and GL were expected to have a focus on unknown and ambiguous terms.

20 annotators were asked to perform only one of the identification tasks, which resulted in five annotations per task. In addition, we asked two annotators to perform all four tasks, to check whether the inter-annotator agreement differs in the two setups. Since the latter annotation setup did not exhibit systematic differences to the original setup, we merged the results of all seven annotations.

Annotation was done using *WebAnno* (Yimam et al., 2013), a general-purpose web-based annotation tool. We allowed annotations of single words, multi-words, and links between terms in case of nonadjacent term constituents. An example of two annotations is shown in figure 1. In addition to the actual annotation, annotators were asked to rate their knowledge about the respective domains. Overall, cooking was rated as best-known domain, with a mean of 6.86 on a scale from 1 (unknown) to 10 (well-known), followed by diy (5.18), chess (4.05) and hunting (1.90).

3 Analyses of Term Identification

In the following, we analyze word forms annotated as terms, across tasks and across domains. As the central means in our analyses, we make use of the *agreement* between annotators. We rely on simple agreement (how many of the 7 annotators per task agreed?), the Jaccard index and the chance-corrected agreement measure Fleiss' κ (Fleiss, 1971). We start with various single-word type-based evaluations in sections 3.1 and 3.2, and then explore multi-words in section 3.3.

3.1 Agreement across Tasks and Domains

Table 1 shows the number of type-based term annotations per task with the highest agreements, i.e. where all annotators (7) or most annotators (6 or 5) agreed. In line with our intuition, the number of identified terms is highest for DS, and lowest for GL, with IND and TR in between.

task	DS	IND	TR	GL
agree = 7 (*all*)	203	66	94	27
agree $\geq$ 6	315	111	173	68
agree $\geq$ 5	400	148	247	117

Table 1: Number of identified terms per task.

This trend is still obvious when including all annotated terms (i.e., all term types annotated by at least one annotator): Figure 2 shows the Jaccard index across tasks and domains, i.e., the intersection of the annotations divided by their union. DS again receives the highest values, GL the lowest. DS and GL thus seem to represent the extremes of the tasks, with DS providing the broadest and GL the narrowest definition of terminology.

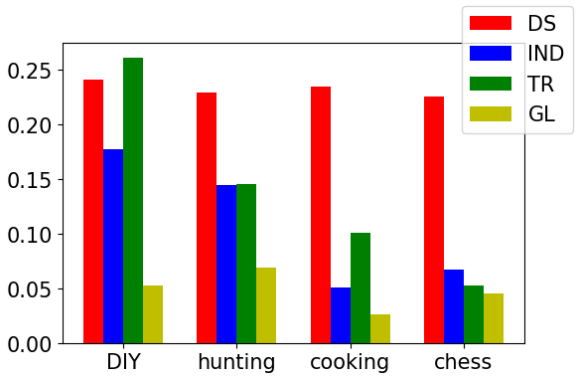

Figure 2: Jaccard index across tasks and domains.

Across the tasks and different scopes of the terms, there is a clear tendency for the same terms to receive high vs. low agreement. This effect is shown in figure 3, where all annotated term types are depicted in a four-dimensional space (x-, y- and z-axis plus the 4th dimension in colour). Each dimension represents one task, the value in each dimension represents the agreement on terms for this task (max. 7). We clearly observe an upward-moving tendency for term agreement across all dimensions, i.e., across the four tasks, annotators (dis-)agreed on the same terms to a similar degree. We conclude that annotators have similar intuitions about a term's domain specificity regardless of the term identification task.

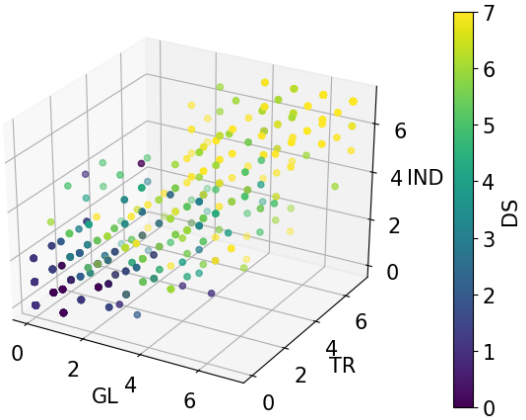

Figure 3: Term agreement across tasks.

Figure 4 depicts the interaction between tasks and domains even more clearly: While Fleiss' κ for DS is in general very high across domains, and also IND and TR are well-agreed upon for diy (and TR for hunting), the κ values for GL are particularly low, and so is IND for cooking and chess, and TR for chess.

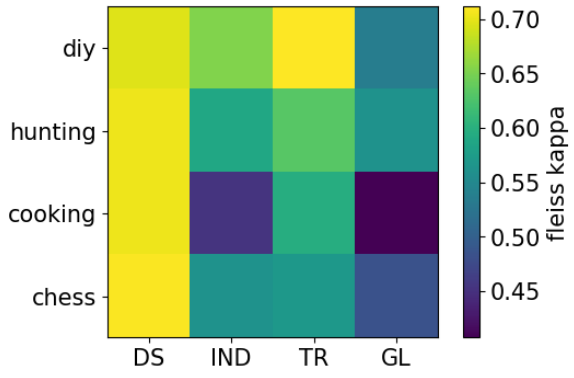

Figure 4: Fleiss' κ across tasks and domains.

3.2 Term Identification across Word Classes

Traditionally, mostly nouns are perceived as terms (Bourigault, 1992; Justeson and Katz, 1995), and consequently annotation and extraction of terms is often restricted to noun phrases (Bernth et al., 2003; Kim et al., 2003). However, according to Estopà (2001) and others, terminology should not be restricted to noun phrases. Figure 5 shows that both views have a point. The figure shows the number of term type annotations for nouns, verbs and adjectives across the 28 annotated datasets (7 annotators $\times$ 4 domains). For example, roughly 300 noun types received a total of 5 term annotations across the four tasks DS, IND, TR and GL.

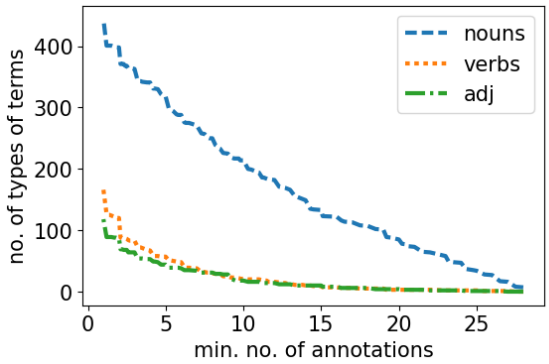

Figure 5: Annotations per part-of-speech.

We can see that in our dataset nouns are indeed preferred by our non-expert annotators. However, when looking at a smaller amount of annotations, the number of annotated verbs and adjectives increases. Looking into the data revealed that 70% and 58% of the annotated verbs and adjectives appear in multi-word terms (MWTs). One reason for this is their participation in annotated activities such as *großes Loch reparieren* ('repair a big hole') or *Eigelb schaumig schlagen* ('beat the egg yolk until fothy').

3.3 Complex Terms and Subterms

The fact that multi-word terms often contain subterms is a distinct attribute, frequently exploited by automatic term extraction methods relying on term constituent phrases for computing a termhood score (Frantzi et al., 1998; Nakagawa and Mori, 2003). In our study, 468 single-word terms, 138 closed compounds and 692 MWT types were annotated across annotators. Since German contains many closed compounds, treating them separately from MWTs (consisting of several separated

words) is especially interesting: a compound term candidate is either annotated completely or not at all. Regarding MWTs, it is possible that only a subterm is annotated. For example, the compound *Rohholz* ('raw wood') cannot be separated, while annotators might mark only *Holz* as subterm of the MWT *rohes Holz*.

Table 2 shows aspects of multi-word and compound term types in relation to their number of annotations (7 annotators × 4 domains, i.e. a maximum of 28 annotations), across tasks and domains. We group the number of annotations into three categories: no concordance (<2), minimum concordance (≥2) and majority concordance (>14). For most MWTs (426), there is no concordance, and only a few MWTs were found in the majority of the 28 annotations (11). Compound terms show the opposite behaviour. Slightly more than half of the compounds (76) were found in the majority of the 28 annotations, while only 6 compounds appeared in only one annotation. Thus, annotators are confident in identifying compound terms, but not MWTs.

no. of annotations	*<2*	*≥2*	*> 14*
MWTs	426	266	11
compounds	6	132	76

Table 2: Annotation of compounds and MWTs.

We then analysed the annotation concordance of complex term components, and their likelihood to represent a subterm, cf. table 3. For that, we extracted all annotated single-word terms (SWTs) which were not also annotated as part of a complex term. We thus obtained the number of annotations for the SWTs only. While for MWTs the proportion of subterms is relatively high across categories (45.83–49.23%), the number of compound subterms is rather low for low-concordance cases (16.67%) and increases radically for higher-concordance cases (up to 40.37%).

no. of annotations	*<2*	*≥2*	*> 14*
MWTs			
% of subterms	49.23	57.40	45.83
∅ annot. on subterms	7.53	7.26	6.0
compounds			
% of subterms	16.67	31.76	40.37
∅ annot. on subterms	1.0	9.59	10.23

Table 3: Annotation of compound and MWT subterms.

Table 3 also illustrates that the average number of annotations per subterm drops for MWTs with an increasing concordance. Compounds, again, behave in the opposite way. Thus, the less confidence there is for an MWT, the more confidence we find in its subterms. For the closed compounds, this effect cannot be perceived.

3.4 Ambiguity

A peculiarity of many terminologies are general-language words with a specialized meaning in one or more domains. For example, the English noun *solution* has a general-language sense as well as domain-specific senses in mathematics and chemistry (Baker, 1988). Ambiguous vocabulary is also present across our domains, e.g., *Fuchsschwanz* ('ripsaw' vs. 'foxtail') in diy and *ansprechen* ('identify game' vs. 'address so.') in hunting.

In order to analyze the identification of ambiguous terms, we first looked up the general-language and domain-specific senses of all hunting and chess terms from our dataset in *Wiktionary*[3], *Duden*[4], and the *Wikipedia* disambiguation pages. We did this for hunting and chess, because only these domains are consistently specified in the sense definitions. We identified 18 terms for hunting and 15 for chess.

Table 4 shows the average agreement on these ambiguous words, across tasks. For example, on average 5.32 annotators out of 7 agreed on the 18 hunting term types in the DS task. The table shows that the average agreement is higher for DS than for the other three tasks.

domain	DS	IND	TR	GL
hunting	5.32	3.74	4.12	3.44
chess	5.08	3.72	3.75	2.93

Table 4: Average agreement on ambiguous words.

We conclude that when it comes to a stricter sense of termhood the domain-specific sense might be perceived by the annotators, but the general-language sense impedes them to accept the same strength of termhood for the ambiguous term as for other, more domain-specific terms.

[3] http://www.wiktionary.org/
[4] https://www.duden.de/

4 Automatic Term Extraction

In a final step, we compared the identification of terms in our dataset against the identification done by state-of-the-art term extraction approaches. We used the hybrid term-candidate extractor for the diy domain described in Schäfer et al. (2015) and Rösiger et al. (2016). After lemmatization and pos-tagging, the system extracts terms with predefined linguistic filters. For term candidate ranking, standard termhood measures are applied, cf. an overview in Schäfer et al. (2015).

Approximately half of our annotated terms were found by the term extractor (due to predefined linguistic patterns for extraction). Based on the measure scores, we applied Spearman's rank-order correlation coefficient ρ (Siegel and Castellan, 1988) to compare against a ranking based on annotator agreement. The best ρ values were 0.51 and 0.44 for two corpus-comparison extraction methods; these are statistically significant ($p < 0.01$).

When inspecting the ranked list, we observed that the term extractors rank compounds and MWTs higher than the laypeople do. Although the automatic extractors only use statistics over the whole word forms, ρ increases when adding subterm scores to compounds and MWTs. This again indicates the importance of subterms within complex terms for an annotator's decision.

5 Conclusion

This paper presented a new dataset of term annotation and a study about term identification by laypeople, across four domains and four task definitions. We found that laypeople generally share a common understanding of termhood and term association with domains, as reflected by inter-annotator agreement. Furthermore,

1. high inter-annotator variance for more specific tasks,
2. little awareness of the degree of termhood of ambiguous terms, and
3. low agreement on multi-word terms with high reliance on subterms

showed that laypeople's judgments deteriorate for specific and potentially unknown terms.

The dataset with the lapeople term annotations is publicly available at `www.ims.uni-stuttgart.de/data/term-annotation-laypeople`.

Acknowledgments

The research was supported by the DFG Collaborative Research Centre SFB 732. We would like to thank Michael Dorna for his helpful comments and reviews.

References

Mihael Arcan. 2017. *Machine Translation of Domain-Specific Expressions within Ontologies and Documents*. Ph.D. thesis, Insight Centre for Data Analytics, National University of Ireland, Galway.

Mihael Arcan, Marco Turchi, Sara Tonelli, and Paul Buitelaar. 2014. Enhancing statistical machine translation with bilingual terminology in a CAT environment. *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)* pages 54–68.

Mona Baker. 1988. Sub-technical vocabulary and the esp teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language* 4(2):91–105.

Gabriel Bernier-Colborne and Patrick Drouin. 2014. Creating a test corpus for term extractors through term annotation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 20(1):50–73.

Arendse Bernth, Michael McCord, and Kara Warburton. 2003. Terminology extraction for global content management. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 9(1):51–69.

Didier Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 15th International Conference on Computational Linguistics*. Nantes, France, pages 977–981.

Fabienne Cap. 2014. *Morphological processing of compounds for statistical machine translation*. Dissertation, Institute for Natural Language Processing (IMS), University of Stuttgart.

M. Teresa Cabré Castellví. 1999. *Terminology: Theory, methods and applications*, volume 1. John Benjamins Publishing.

Rosa Estopà. 2001. Les unités de signification spécialisées: élargissant l'objet du travail en terminologie. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 7(2):217–237.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

Katerina T. Frantzi, Sophia Ananiadou, and Jun-ichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*. London, UK, pages 585–604.

Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, and Herbert Ernst Wiegand. 2007. Dictionaries. An international encyclopedia of lexicography. *Supplementary Volume: Recent Developments with special focus on Computational Lexicography* .

Anna Hätty, Simon Tannert, and Ulrich Heid. 2017. Creating a gold standard corpus for terminological annotation from online forum data. In *Proceedings of the EACL Workshop on Language, Ontology, Terminology and Knowledge Structures*.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1(1):9–27.

J.-D. Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. GENIA corpus – A semantically annotated corpus for bio-textmining. *Bioinformatics* 19(1):180–182.

Hiroshi Nakagawa and Tatsunori Mori. 2003. Automatic term recognition based on statistics of compound nouns and their components. *Terminology* 9(2):201–219.

Thorsten Roelcke. 1999. *Fachsprachen*. Grundlagen der Germanistik. Erich Schmidt Verlag.

Ina Rösiger, Julia Bettinger, Johannes Schäfer, Michael Dorna, and Ulrich Heid. 2016. Acquisition of semantic relations between terms: How far can we get with standard NLP tools? In *Proceedings of the 5th International Workshop on Computational Terminology*.

Johannes Schäfer, Ina Rösiger, Ulrich Heid, and Michael Dorna. 2015. Evaluating noise reduction strategies for terminology extraction. In *Proceedings of TIA 2015*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, pages 44–49.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Boston, MA.

Louis Trimble. 1985. *English for Science and Technology: A Discourse Approach*. Cambridge University Press.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria, pages 1–6.

Behrang Qasemi Zadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. Portoroz, Slovenia.