Predicting Degrees of Technicality in Automatic Terminology Extraction

Anna Hätty[1,2], Dominik Schlechtweg[2], Michael Dorna[1], Sabine Schulte im Walde[2]
[1]Robert Bosch GmbH, Corporate Research, Renningen, Germany
[2]Institute for Natural Language Processing, University of Stuttgart, Germany
`{anna.haetty, michael.dorna}@de.bosch.com,`
`{schlecdk,schulte}@ims.uni-stuttgart.de`

Abstract

While automatic term extraction is a well-researched area, computational approaches to distinguish between degrees of technicality are still understudied. We semi-automatically create a German gold standard of technicality across four domains, and illustrate the impact of a web-crawled general-language corpus on predicting technicality. When defining a classification approach that combines general-language and domain-specific word embeddings, we go beyond previous work and align vector spaces to gain comparative embeddings. We suggest two novel models to exploit general- vs. domain-specific comparisons: a simple neural network model with pre-computed comparative-embedding information as input, and a multi-channel model computing the comparison internally. Both models outperform previous approaches, with the multi-channel model performing best.

1 Introduction

Automatic term extraction, i.e. the task of extracting linguistic expressions characteristic to a specialized domain, is a long-researched field within natural language processing. Assessing the technicality of the extracted terms, however, is still a niche within this area: technicality refers to the *degree* to which a term is specialized and exclusively used by experts in a domain. Up to date, studies on term technicality are mostly restricted to medical terminology and relate to the communication between doctors and patients. Especially in times of growing amounts of domain-specific websites with both lay and expert users (e.g. DIY 'do-it-yourself' communities, such as `1-2-do.com`), the communication between experts and lays becomes increasingly important across all specialized domains. Furthermore, term technicality prediction is important for a range of tasks such as automatic thesaurus

creation, assessing text specialization, and domain knowledge acquisition. Above all, predicting technicality can be considered a more fine-grained and expressive form of terminology extraction.

In this work, we first semi-automatically collect German specialized domain corpora to create a gold standard of term technicality across four domains: automotive, cooking, hunting and DIY. Based on a qualitative analysis of terminological phenomena and variants of ambiguity across domain-specific and general-language corpora, we then suggest two methods to explicitly integrate not only vector space model representations derived from the corpora, but also comparisons across the vector spaces. In a first approach, we enrich the combined general-language and domain-specific word embeddings with a difference vector as input for a classification system. In a second approach we design a multi-channel feed-forward neural network with a Siamese network component to represent the vector comparison internally.

2 Related Work

Existing studies on technicality predominantly focus on levels of familiarity or difficulty of terminology in medical, biomedical or health domains. Term familiarity refers to a user's subjective understanding of term technicality. These studies typically rely on classical readability features such as frequency, term length, syllable count, the Dale-Chall readability formula and affixes (Zeng et al., 2005; Zeng-Treitler et al., 2008; Grabar et al., 2014; Vinod Vydiswaran et al., 2014). They further make use of domain-specific terminology attributes such as neo-classical word components, given that medical terminology is strongly influenced by Greek and Latin (Deléger and Zweigenbaum, 2009; Bouamor et al., 2016). Besides the feature specification, the majority of studies exploits contrastive approaches.

Contrastive approaches compare a term's distribution in a domain and a reference corpus, for example a general-language corpus. Furthermore, for technicality prediction, often expert (medical) texts are compared against reference lay texts. Only a small number of studies relies on context-based approaches, e.g. Zeng-Treitler et al. (2008) use a contextual network; Bouamor et al. (2016) exploit language models; Pérez (2016) compares collocation networks.

For standard term extraction, contrastive techniques represent one of the main strands of methodologies, by comparing a term candidate's frequencies in a domain-specific and a general-language corpus (Ahmad et al., 1994; Rayson and Garside, 2000; Drouin, 2003; Kit and Liu, 2008; Bonin et al., 2010; Kochetkova, 2015; Lopes et al., 2016; Mykowiecka et al., 2018, i.a.). Recent approaches use word embeddings trained separately on contrastive corpora; e.g. Amjadian et al. (2016, 2018) concatenate general and domain-specific word embeddings and use them as input for classifiers, such as a multilayer perceptron. Similarly, Hazem and Morin (2017) and Liu et al. (2018) apply such a concatenation to represent a term in one language, as data enrichment pre-step for bilingual terminology extraction.

In sum, approaches using contrastive corpora are popular in both automatic term extraction and term technicality prediction studies. The few approaches that use word embeddings as basis for a contrastive approach separately train word embeddings on general-language and domain corpora. In our work, we extend these methodologies by aligning vector spaces in order to more adequately represent meaning variation across corpora.

3 Definition of Technicality

According to Ha and Hyland (2017), there is no consensus among researchers about what exactly defines technicality. They provide an overview of what characterizes technical vocabulary, and observe two main categories. On the one hand, technical terms often exhibit a narrow range of senses specific to the domain. They are only understood by a limited set of people, because they require domain knowledge. On the other hand, there are terms which are also frequently used in general language. These terms are ambiguous: they carry specialized meanings in a particular domain which are different to the general-language meanings.

Corpus sizes	Preprocessed	Lemma:POS
Cooking	4.3 M	2.5 M
Automotive	4.9 M	2.3 M
DIY	4.0 M	2.1 M
Hunting	0.7 M	0.3 M
SdeWaC	778 M	326 M

Table 1: Sizes of corpora. "Preprocessed" refers to the lemmatized corpus without punctuation, "Lemma:POS" to the version reduced to content words.

As Ha and Hyland (2017), we see technicality as a continuum. In the course of this paper, we adopt a simplified handling and distinguish between three broad classes of technicality: technical terms, basic terms and non-terms.

4 Data and Gold Standard Creation

Data. We collect German texts for four domains: automotive, cooking, DIY and hunting. Besides including technical handbooks, we crawl topic-specific data from Wikipedia[1] and similar resources such as cooking recipes from cooking homepages (e.g. `kochwiki.de`), and car repair and DIY instructions from `wikihow.de`. As general-language reference corpus, we use *SdeWaC* (Faaß and Eckart, 2013), a cleaned version of the web-crawled corpus *deWaC* (Baroni et al., 2009). All corpora are lemmatized and POS-tagged with the *TreeTagger* (Schmid, 1995), and reduced to content words (nouns, verbs and adjectives). We follow the preprocessing steps described in Schlechtweg et al. (2019) that led to the best results in that study. The corpus sizes are shown in Table 1.

Gold Standard. We select all words as term candidates with a minimum frequency of 10 in both the domain corpus and SdeWaC. The gold standard thus contains both simple and complex terms, the latter in the form of closed compounds. We did not extract multi-word terms other than closed compounds because we would have needed specific procedures to identify them (e.g. by chunking or by using association measures to identify valid collocations). Even more importantly, multi-word expressions are prone to variation (e.g. one could say 'wood drill' or 'drill for wood') and it is likely to not find all variants in the glossaries and other resources we use to create the gold standard.

---

[1]When using Wikipedia, we relied on grouping categories such as the category 'automotive' (https://de.wikipedia.org/wiki/Kategorie:Kraftfahrzeugtechnik).

Instead of relying on labour-intensive human annotations, we determine the technicality labels semi-automatically. First, we collect domain-specific glossaries for each domain, i.e. textual glosses and specialized terms with their meanings[2]. These glossaries contain terms which require domain knowledge (especially if they are ambiguous) and thus need to be explained to a lay person, i.e. they contain technical terms. Secondly, we collect thematic basic vocabulary lists (from thematic base vocabulary books, thematic vocabulary training lists for foreign apprentices, etc.). These lists contain the basic terminology of a domain, with a low level of technicality. Finally, we collect indices and tables of contents of domain-specific handbooks, which include all kinds of terminological vocabulary[3]. We label the data as follows:

1. **technical term**: a word is contained in a glossary, but not in a basic vocabulary list

2. **basic term**: a word is contained in a basic vocabulary list, but not in a glossary

3. **non-term**: all other words, which do not overlap more than 4 characters with any term in the glossaries, the basic vocabulary lists, the indices or the table of contents

The resulting sizes of the gold standards per domain are presented in Table 2. Overall, our semi-automatic labeling method leads to 1,690 technical terms, 1,525 terms and 10,956 non-terms, a total of 14,171 term candidates. To evaluate the quality of the gold standard, we randomly extract 30 words per domain and per system-assigned label (which leads to a total of $30 \times 4 \times 3 = 360$ words in total). Together with three random context sentences, three annotators (including one of the authors) rated the labeling. We obtain an average Cohen's κ inter-annotator agreement of 0.50 and an average agreement with the gold standard of 0.47. This corresponds to "moderate" agreement, which we judge as sufficient for our gold standard, given that agreement in term annotation is considered a difficult task (Terryn et al., 2019).

	Cook.	Hunt.	Auto.	DIY
Tech. Terms	384	250	706	350
Basic Terms	853	186	236	250
Non-Terms	853	1,176	5,010	2,962
Total	3,045	1,612	5,952	3,562

Table 2: Size of gold standard.

Qualitative Analysis We perform an in-depth analysis of our four domain corpora to identify the range of term phenomena and variants of ambiguity within and across general and domain-specific data, to motivate and apply an appropriate model.

The automotive domain contains many compounds (such as *Antriebsschlupfregelung* 'traction slip control') and English words (*Frontairbags*). In the cooking and DIY corpora we find many complex verbs (such as *entgraten* 'deburr' for DIY and *abbinden* 'thicken (a sauce)' for cooking). Ambiguous terminology is an outstanding characteristic of the hunting domain, which contains many ambiguous expressions completely unknown by lay people, such as *Licht* 'light' as term for the eyes of game. With all those variations, it seems likely that surface form features will not be useful in a prediction task. Furthermore, frequency-based features might not be useful due to the high amount of ambiguity.

Regarding levels of technicality, we find technical terms that seem to be rather unambiguous and have a very restricted usage, such as *blanchieren* 'blanch' for cooking, which often co-occurs with *Salzwasser* 'salted water' in the domain-specific context sentences. Surprisingly, we find very similar domain-specific contexts in the general-language corpus, where we would not expect them. Since the general-language corpus is web-crawled, it obviously contains a certain amount of domain-specific texts as well; especially if a highly technical term is not ambiguous, the general-language corpus contains only such contexts. Consequently, the general-language and domain-specific contexts are maximally similar in these cases. In contrast, we assume that the contexts will vary more strongly for basic terms, and for non-terms we do not expect to find domain-specific sentences in the general-language corpus at all.

The picture is different for ambiguous terminology, where sense distributions vary across corpora. For example, for the hunting term *Licht* 'light/eyes of game' we both find general and domain-specific meanings in the domain corpus; for the cooking term *Zauberstab* 'wand/hand blender' senses seem

to be largely disjunctive across the corpora. Example sentences for this phenomenon are given in Table 3 for illustration.

Based on these observations, we suggest an approach by Amjadian et al. (2016, 2018) as basis to detect degrees of technicality, since both general-language and domain-specific word embeddings will encode termhood attributes. On top of that, we hypothesize that a comparison of the word vectors represents valuable information for a prediction system.

5 Models

Baselines As baseline, we use a decision tree classifier (DT) with three standard features commonly used for term familiarity prediction: frequency (corpus-size normalized), word length and character n-grams. Further, we implement the approach by Amjadian et al. (2016, 2018) using a Multilayer Perceptron (MLP) and the concatenation of general-language word embeddings (GEN) and domain-specific word embeddings (SPEC) of a term candidate as input (MLP, GEN⊕SPEC), in comparison to using only one of the embeddings. We learn two separate word2vec SGNS vector spaces (Mikolov et al., 2013) for GEN and SPEC.

Centering and Batch Normalization Across neural models we apply batch normalization (Ioffe and Szegedy, 2015), which normalizes the output of a preceding activation layer by subtracting the batch mean and then dividing by the batch standard deviation. This reduces the effect of inhomogeneous input data, in our case the different domain corpora. We further length-normalize and apply element-wise column mean-centering to the embeddings, which has proven to be beneficial as preprocessing step for rotational alignment of vector spaces (Artetxe et al., 2016; Schlechtweg et al., 2019) and as a general post-processing step for word embeddings (Mu and Viswanath, 2018).

Note that the reason for the beneficial effect of centering is still unclear. Artetxe et al. (2016) provide an intuitive explanation that centering moves randomly similar embeddings further apart, while Mu and Viswanath (2018) consider centering as an operation making vectors "more isotropic", i.e., more uniformly distributed across the directions in the space.

Comparative Embeddings and Multi-Channel Model Simple vector concatenation does not in-

corporate any kind of comparison of the embeddings. We thus suggest two novel models to exploit general- vs. domain-specific comparisons: *Comparative Embeddings* (MLP, CON⊕DIFF) use an MLP classifier and add a difference vector to the input vector concatenation GEN⊕SPEC. Since the word embeddings were trained separately on different corpora, this model requires an alignment of the vector spaces. We use a state-of-the-art alignment method (Artetxe et al., 2016; Hazem and Morin, 2017), where the best rotation GW of a vector space G onto a vector space S is determined, with the rotation matrix W. W is computed as $W = UV^T$, with U and V retrieved from Singular Value Decomposition $S^TG = U\Sigma V^T$ (Schönemann, 1966). After the alignment, unit length is applied again (since the vectors are not unit length after alignment anymore) and the absolute difference vector (DIFF) is computed. The concatenation vector GEN ⊕ SPEC ⊕ DIFF is then taken as input to the model.

As our second model, we use a *Multi-Channel Feed-Forward Neural Network* (MULTI-CHANNEL). The network takes as input the unaligned GEN and SPEC vectors, and processes each GEN and SPEC in a different channel. The third channel is a variant of a *Siamese network* (Chopra et al., 2005), a dual-channel network with shared weights. Both GEN and SPEC are processed through the shared weight layer, in order to map them onto the same space. Then the element-wise absolute difference is computed, and the output of all three channels is concatenated. The network is defined as:

$$h_1 = \sigma_1(W_1 * E(x_1) + b_1)$$
$$h_2 = \sigma_2(W_2 * E(x_2) + b_2)$$
$$h_{3a} = \sigma_3(W_3 * E(x_1) + b_3)$$
$$h_{3b} = \sigma_3(W_3 * E(x_2) + b_3)$$
$$d = |h_{3a} - h_{3b}|, \ d \in R^l$$
$$c = h_1||h_2||d, \ c \in R^{3l}$$
$$p = softmax(c)$$

where x is a term candidate, and $E(x)$ is the embedding layer, a function $E : x_i \rightarrow z_i$ that maps the word x_i onto its corresponding 300-dimensional vector z_i. W denotes the weight matrices, b the bias, σ the activation functions, and l denotes the sizes of the hidden layers.

General-language corpus	Domain-specific corpus
Ich denke, mit **Zauberstab** kann man leichter zaubern.	1 Mixgerät, Handrührer mit Mixstab oder **Zauberstab** mit Schüssel
Nicht vergessen soll er bitte seinen **Zauberstab** und es bleibt ihm freigestellt, ob er eine Eule, eine Katze oder eine Kröte mitbringt.	Die Sauce abermals erhitzen, die Butter mit der Stopfleber zugeben und die Sauce mit einem **Zauberstab** schaumig aufmixen.
Ich verließ die Bank und wanderte mit dem Blick gebannt auf den Mond, taumelnd, wie hypnotisiert, dem **Licht** entgegen.	**Lichter** ist die Bezeichnung für die Augen, die Ohren werden auch Lauscher genannt.
Mit Betten, **Licht** und einem Tisch.	Auch bei schwachem **Licht** können sie noch sehr gut sehen.

Table 3: Example context sentences for the ambiguous terms *Zauberstab* (cooking, upper table) and *Licht* (hunting, lower table). The sentence with a lime green background contains the target term in its general-language sense.

Training We use SMOTE subsampling (Chawla et al., 2002) and train our network to minimize the cross-entropy loss, using back-propagation with stochastic gradient descent. We perform a randomized search for hyperparameter optimization for each model, i.e. subsampling parameter combinations. We test with the following parameters: hidden layers, epochs and batch size with values between 16 and 64, learning rate between 0.001 and 0.3, momentum between 0.0 and 0.9, and *tanh* and *rectified linear unit (ReLU)* as activation functions. To initialize the weights of the embedding layer, we use word2vec SGNS trained with a window size of 2, negative sampling with k=1 and subsampling with a threshold of $t = 0.001$. These parameter settings obtained the best results in our recent study on terminological meaning shifts (Schlechtweg et al., 2019). We do not train embedding layer parameters to maintain the original word meaning. Due to the relatively small size of the training data, we use 5-fold cross-validation for training.

6 Results

We use Macro-Precision, Recall and F1-Score for evaluation, to put more weight on the correctness of the smaller classes *Base Term* and *Technical Term*. The experiment results are shown in Table 4.

The multi-layer perceptron (MLP) results outperform the decision-tree (DT) baseline with standard term familiarity prediction features. Using only a general-language vector GEN for classification performs better than using only a domain-specific vector SPEC, and the concatenation of both

Method	P	R	F1
DT, basic features	0.56	0.58	0.57 (−)
MLP, SPEC	0.65	0.79	0.69 (0.62)
MLP, GEN	0.68	0.82	0.73 (0.72)
MLP, GEN⊕SPEC	0.76	0.89	0.81 (0.76)
MLP, CON⊕DIFF	0.84	0.94	0.88 (0.88)
MULTI-CHANNEL	**0.86**	**0.94**	**0.89 (0.85)**

Table 4: Macro-Precision (P), Recall (R) and F1-Score results. The main results apply centering and batch normalization; results without centering are in brackets.

(GEN⊕SPEC) performs better than each of them individually. This is most likely due to more training data and having both domain-specific and general-language parts in the general-language corpus.

The models integrating a notion of vector comparison perform best, with the multi-channel network achieving slightly better results than the MLP comparative embeddings. Centering improves all but one results; i.e., it has an overall beneficial effect for our task.

7 Conclusion

We semi-automatically created the first large-scale gold standard for technicality prediction across domains and proposed two novel neural network models to fine-tune automatic terminology extraction by distinguishing between degrees of technicality. The models integrate general- vs. domain-specific word embedding information in different ways. An adapted Siamese multi-channel network model performed best, and centering has an overall beneficial effect on pre-processing the vector spaces.

References

Khurshid Ahmad, Andrea Davies, Heather Fulford, and Margaret Rogers. 1994. What is a term? The semi-automatic extraction of terms from text. *Translation Studies: An Interdiscipline. Selected papers from the Translation Studies Congress,Vienna, 1992*, 2:267–278.

Ehsan Amjadian, Diana Inkpen, T Sima Paribakht, and Farahnaz Faez. 2016. Local-global vectors to improve unigram terminology extraction. In *Proceedings of the 5th International Workshop on Computational Terminology*, pages 2–11, Osaka, Japan.

Ehsan Amjadian, Diana Inkpen, T Sima Paribakht, and Farahnaz Faez. 2018. Distributed specificity for automatic terminology extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 24(1):23–40.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Armin Blass and Wolf Friederich. 1974. *Englischer Wortschatz in Sachgruppen*. 2077. Hueber Verlag.

Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2010. A contrastive approach to multi-word term extraction from domain corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 19—21, Valletta, Malta.

Dhouha Bouamor, Leonardo Campillos Llanos, Anne-Laure Ligozat, Sophie Rosset, and Pierre Zweigenbaum. 2016. Transfer-based learning-to-rank assessment of medical term technicality. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Paris, France.

Nitesh Chawla, Kevin Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research (JAIR)*, 16:321–357.

Sumit Chopra, Raia Hadsell, and Yann Lecun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of Computer Vision and Pattern Recognition Conference*, pages 539–546.

Louise Deléger and Pierre Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pages 2–10.

Karl-Heinz Dietsche, Konrad Reif, et al. 2019. *Kraft-fahrtechnisches Taschenbuch*. Robert Bosch GmbH (Eds.), Springer Vieweg, Wiesbaden.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(1):99–115.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – A corpus of parsable sentences from the web. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer, Berlin Heidelberg.

Natalia Grabar, Thierry Hamon, and Dany Amiot. 2014. Automatic diagnosis of understanding of medical words. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–20, Gothenburg, Sweden.

Althea Ying Ho Ha and Ken Hyland. 2017. What is technicality? A technicality analysis model for EAP vocabulary. *Journal of English for Academic Purposes*, 28:35–49.

Amir Hazem and Emmanuel Morin. 2017. Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 685–693, Taipei, Taiwan.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456, Lille, France.

Chunyu Kit and Xiaoyue Liu. 2008. Measuring monoword termhood by rank difference via corpus comparison. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 14(2):204–229.

Natalia A. Kochetkova. 2015. A method for extracting technical terms using the modified weirdness measure. *Automatic Documentation and Mathematical Linguistics*, 49(3):89–95.

Jingshu Liu, Emmanuel Morin, and Peña Saldarriaga. 2018. Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2855–2866, Santa Fe, New Mexico, USA.

Lucelene Lopes, Paulo Fernandes, and Renata Vieira. 2016. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf. *Knowledge-Based Systems*, 97:237–249.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119, Lake Tahoe, Nevada, USA. Curran Associates Inc.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Piotr Rychlik. 2018. Recognition of irrelevant phrases in automatically extracted lists of domain terms. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 24(1):66–90.

María José Marín Pérez. 2016. Measuring the degree of specialisation of sub-technical legal terms through corpus comparison. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 22(1):80–102.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

Peter H Schönemann. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10.

M. Schroder. 2006. Franz Dornseiff: Der deutsche Wortschatz nach Sachgruppen. 8., vollständig neu bearbeitete und mit einem vollständigen alphabetischen Zugriffsregister versehene Auflage von Uwe Quasthoff. *Deutsch als Fremdsprache*, 43(1):53.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2019. In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, pages 1–34.

V.G. Vinod Vydiswaran, Qiaozhu Mei, David A. Hanauer, and Kai Zheng. 2014. Mining consumer health vocabulary from community-generated text. In *Proceedings of the AMIA Annual Symposium*, volume 2014, pages 1150–1159.

Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A text corpora-based estimation of the familiarity of health terminology. In *International Symposium on Biological and Medical Data Analysis*, pages 184–192. Springer, Berlin/Heidelberg.

Qing Zeng-Treitler, Sergey Goryachev, Tony Tse, Alla Keselman, and Aziz Boxwala. 2008. Estimating consumer familiarity with health terminology: A context-based approach. *Journal of the American Medical Informatics Association*, 15(3):349–356.