SURel: A Gold Standard for Incorporating Meaning Shifts into Term Extraction

Anna Hätty[1,2], Dominik Schlechtweg[2], Sabine Schulte im Walde[2]
[1]Robert Bosch GmbH
[2]Institute for Natural Language Proessing (IMS), University of Stuttgart
anna.haetty@de.bosch.com, {schlecdk,schulte}@ims.uni-stuttgart.de

Abstract

We introduce SURel, a novel dataset for German with human-annotated meaning shifts between general-language and domain-specific contexts. We show that meaning shifts of term candidates cause errors in term extraction, and demonstrate that the SURel annotation reflects these errors. Furthermore, we illustrate that SURel enables us to assess optimisations of term extraction techniques when incorporating meaning shifts.

1 Introduction

Domain-specific terms often undergo meaning shifts from general-language use to their respective domain-specific language use. For example, the German noun *Schnee* predominantly means 'snow' in its general-language usage, and 'beaten egg whites' in the cooking domain. Terms with these characteristics are referred to as *sub-technical terms* and pose a problem for term extraction: Their hybrid character makes it hard for humans to rank them along with unambiguous terms, and hard for computational models to classify them as terms, because of the strong bias towards their general-language meanings.

In this study, we present SURel (**S**ynchronic **U**sage **R**elatedness), a novel dataset for meaning shifts from general to domain-specific language, based on human annotations on the degrees of semantic relatedness between contexts of term candidates. We illustrate that SURel reflects the error that is commonly made by term extraction measures for sub-technical terms when relying on a general-language reference corpus. In a first experiment, we predict the meaning shift automatically and use SURel for evaluation. We then incorporate the model's prediction as a factor into an established term extraction measure, to correct the error in termhood prediction caused by meaning shifts.

2 Meaning Shifts in Terminology

Sub-Technical Terms Terms are linguistic units that characterize specialized domains (Kageura and Umino, 1996), thus representing opposite extremes of words that are not specific to a domain (Sager, 1990). *Sub-technical terms* (Cowan, 1974; Trimble, 1985; Baker, 1988; Chung and Nation, 2003; Pérez, 2016) occupy intermediary positions on the continuum, because they undergo meaning shifts from general to domain-specific language usage. Baker (1988) distinguishes two types of sub-technical terms with general-language usage: words with a restricted domain-specific meaning (e.g., *effective* means 'take effect' in biology), and words with a complete meaning shift (e.g., *bug* in computer science).

Sub-technical terms are a major problem for term extraction measures which often operate on the word type rather than the word sense level. Pérez (2016) provides empirical evidence that 50% of legal terminology is represented by sub-technical terms. Lay people often do not even notice their terminological character due to their predominant general-language use (Hätty and Schulte im Walde, 2018).

Term Extraction Techniques One of the main strands of term extraction methodologies are *contrastive* techniques, which compare a term candidate 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.). For these methods sub-technical terms are problematic, because their meanings are biased towards their general-language use. An illustration is given in Figure 1.

Contrastive term extraction measures are usually designed to identify terms with meaning *stability*, i.e., the meaning in a domain-specific cor-

*Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 1–8
Minneapolis, June 6–7, 2019.

pus is the same as the meaning in a general-language corpus. If a term candidate undergoes a meaning shift, either a meaning *reduction* takes place, i.e., only a subset of the general-language meanings occurs in a domain-specific corpus, or we find a complete meaning *change*. Both reduction and change cause errors in the term extraction results, which are stronger for meaning change in comparison to meaning reduction.

It is evident that there are occurrences of senses in the general-language corpus which should not be considered as term meanings (see hatchings).

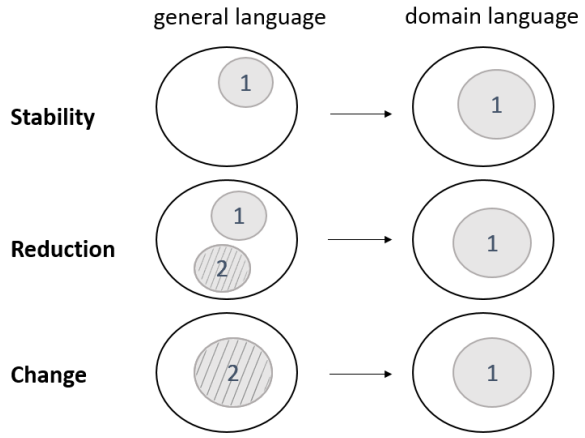

Figure 1: Influence of meaning shifts on a term's sense distributions across languages.

With very few exceptions, sub-technical terms are not explicitly addressed by contrastive measures. Drouin (2004) mentions in his qualitative analysis that some polysemous terms are not found by his extraction system. Menon and Mukundan (2010) and Pérez (2016) do explicitly tackle the extraction of sub-technical terms. Their systems rely on a term candidate's collocation frequencies in a domain and a general reference corpus. But due to the lack of a gold standard, they only perform a qualitative analysis.

This is where our work comes into play: sub-technical terms could be extracted in the same way as terms, if only the corresponding meanings were taken into account when comparing general-language and domain-specific uses. Our novel dataset SURel captures meaning shifts of term candidates and thus serves as a gold standard for the strength of the expected error produced by contrastive term extraction techniques when applied to sub-technical terms.

3 The Dataset: SURel[1]

Dataset Creation SURel was created analogously to DURel (Schlechtweg et al., 2018), a dataset for meaning shifts across time. Our novel dataset comprises a manual annotation of meaning relatedness between uses of target words in a general-language and a domain-specific corpus. The strength of relatedness between uses defines whether the meanings of a word are related or differ, thus indicating if a meaning shift took place.

As general-language corpus (GEN) we subsampled SdeWaC (Faaß and Eckart, 2013), a cleaned version of the web corpus DEWAC (Baroni et al., 2009). As domain-specific corpus (SPEC), we crawled cooking-related texts from several categories (recipes, ingredients, cookware, cooking techniques) from the German cooking recipe websites *kochwiki.de* and *Wikibooks Kochbuch*[2]. The reduced SdeWaC contains ≈126 million words, SPEC contains ≈1.3 million words.

We selected 22 target words which occurred in both GEN and SPEC, and which we expected to exhibit different degrees of domain-specific meaning shift. For each target word we randomly sampled 20 use pairs (i.e., combinations of two contexts) from GEN, SPEC and across both, a total of 60 use pairs per word and 1,320 use pairs overall. Four native speakers annotated the use pairs on a scale from 1 (unrelated meanings) to 4 (identical meanings), reaching a strong mean pairwise agreement of ρ =0.88. The ranking of the 22 target words by their average strength of relatedness between general-language and domain-specific uses is shown in Figure 2. On the left are target words with highly related meanings in GEN and SPEC; on the right are words with strongly different meanings.[3]

Dataset Analysis In the following, we analyse the meaning relatedness of use pairs within and across GEN and SPEC. Figure 3 shows examples of annotations that nicely correspond to cases of meaning *stability, reduction* and *change*, respectively. The y-axes show how often the use pairs were rated as 1–4. In Figure 3 top left we find *Schnittlauch* 'chive' with strongly related meanings within and across GEN and SPEC, thus indicating meaning stability. Top right, we find

[1]The dataset is available at www.ims.uni-stuttgart.de/data/surel.

[2]de.wikibooks.org/wiki/Kochbuch

[3]Find an overview of the dataset in the Appendix.

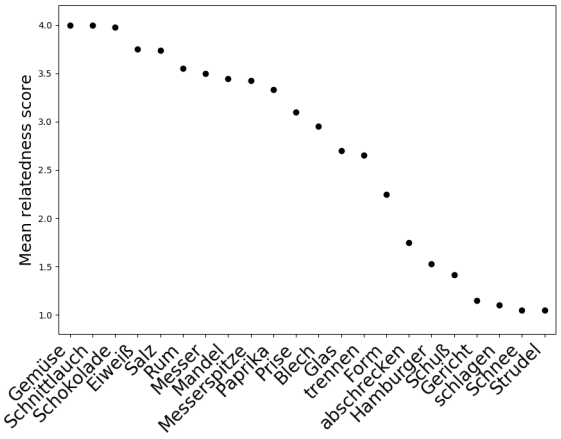

Figure 2: Ranking of target words by average strength of meaning relatedness between GEN and SPEC.

Messer 'knife' with more related meanings in SPEC than in GEN, and even less strongly related meanings across GEN and SPEC, thus indicating meaning reduction. In Figure 3 at the bottom we find *Schnee* 'snow'/'beaten egg whites' with strongly related meanings within GEN and also within SPEC but very different meanings when comparing GEN and SPEC uses, thus indicating a meaning shift. The three examples are taken from the two extremes and a mid position in Figure 2.

4 Incorporating Meaning Shifts into Automatic Term Extraction

After illustrating that the relatedness scores in SURel reflect degrees of meaning shifts from general to domain-specific language usage, the current section demonstrates that (a) a standard measure for automatic term extraction does not capture variants of meaning shifts, and (b) we can utilise SURel to modify existing measures to incorporate meaning shifts into termhood prediction.

A Standard Term Extraction Measure We selected one of the simplest standard contrastive term extraction measures, the *Weirdness Ratio* (WEIRD) (Ahmad et al., 1994), which is still commonly used or adapted (Moreno-Ortiz and Fernández-Cruz, 2015; Cram and Daille, 2016; Roesiger et al., 2016; Hätty et al., 2017, i.a.). It encompasses just the basic ingredients for termhood prediction, a comparison of word frequencies in relation to corpus sizes:

$$\text{WEIRD}(x) = \frac{f_{spec}(x)/s_{spec}}{f_{gen}(x)/s_{gen}},$$

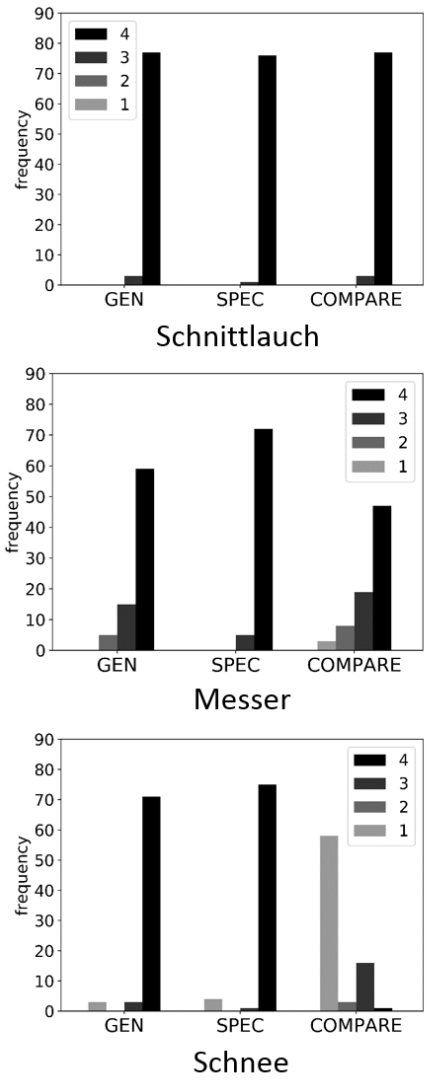

Figure 3: Examples indicating meaning stability (top), meaning reduction (centre) and meaning change (bottom). COMPARE denotes cross-corpora relatedness (cf. Schlechtweg et al., 2018).

where f_{spec} and f_{gen} correspond to the frequencies of a term candidate x in a general and a domain-specific corpus, and s_{spec} and s_{gen} are the respective sizes of the corpora.[4]

The left panel in Figure 4 shows the ranking of the SURel target words after computing their WEIRD scores, with decreasing termhood scores for targets from left to right. The figure clearly illustrates that WEIRD ranks the targets words with strongest meaning shifts in SURel lowest, independently of their termhood: targets with high SURel scores are ranked as most terminological by WEIRD and occupy the first ranks (*Messerspitze, Eiweiß*, ...), and targets with low SURel

[4]We use versions of our corpora which are limited to content words to be consistent with following experiments.

scores are ranked as the least terminological ones and occupy the last ranks (. . . , *Form, schlagen*).

To further investigate this bias, we looked up the SURel targets in (a) Wiktionary and Wikipedia, (b) the German dictionary Duden and (c) popular German translation dictionaries (Langenscheidt and PONS). If a word was assigned a cooking or gastronomy tag in any of these resources, we categorised it as a domain term. In this way, ten of our targets[5] were categorised as terms; seven of them are among the ten most non-terminologically ranked targets by WEIRD. This confirms that termhood predictions by WEIRD as a representative of contrastive termhood measures are strongly influenced by terminological meaning shifts.

Although the influence of meaning shifts might not be equally evident in other term extraction measures as in our simple example measure WEIRD, any other measure heavily relying on a general-language word frequency distribution will to some extent be negatively influenced by terminological meaning shifts. Consequently, we need to correct the bias caused by meaning shifts. In the following, we show that we can use SURel to assess factors that potentially reduce the bias.

Correcting the Meaning Shift For automatically predicting meaning shifts we rely on a state-of-the-art model for diachronic meaning change (Hamilton et al., 2016). We learn two separate word2vec SGNS vector spaces for GEN and SPEC. In order to compare the target vectors across spaces the spaces are aligned, i.e., the best rotation of one vector space onto the other is computed. This corresponds to the solution of the orthogonal Procrustes problem (Schönemann, 1966). If G and S are the matrices for the general and the specific vector spaces, then we rotate G by GW where $W = UV^T$, with U and V retrieved from the singular value decomposition $S^TG = U\Sigma V^T$. Following standard practice we then length-normalize and mean-center G and S in a pre-processing step (Artetxe et al., 2017). After the alignment, cosine similarity between the vectors of the same word in both spaces is computed. The cosine score of the two vectors of a word w predicts the strength of meaning change of w between GEN and SPEC, ranging from 0 (complete change) to 1 (stability).[6]

As input for the model, we use POS-tagged versions of our corpora, keeping only content words.

Evaluating the output of the model on the SURel dataset, we reach a Spearman's rank-order correlation coefficient of ρ=0.866 between the model's change predictions and SURel meaning-shift ranks. Inspecting the nearest neighbors (NNs) of our target words in Figure 3 confirms the ability of the model to predict strengths of meaning shifts. For example, the NNs for *Schnee* change completely (from *mud, leaves, foggy* in the GEN space to *egg whites, foamy, beat* in the SPEC space), while for *Schnittlauch* all nearest neighbors in both spaces are cooking-related.

Finally, to correct WEIRD for the meaning-shift error, we incorporate the model's predictions of meaning change into the WEIRD formula, where $\alpha(x)$ corresponds to the model's predicted strength of meaning change for word x:

$$\text{WEIRD}_{MOD}(x) = \frac{f_{spec}(x)/s_{spec}}{(\alpha(x) \cdot f_{gen}(x))/s_{gen}}.$$

The right panel in Figure 4 shows the ranking of the SURel target words based on their WEIRD_{MOD} scores, again with decreasing termhood scores for targets from left to right. The plot clearly shows that WEIRD_{MOD} improves over WEIRD regarding the negative bias for meaning-shifted targets: now shifted target words do not gather in one part of the plot but occur across ranks. While WEIRD only reaches an average precision of 0.45, WEIRD_{MOD} reaches an average precision of 0.59.

In the same way as we incorporated the Hamilton measure of semantic change into WEIRD, we could rely on other contrastive term extraction techniques and incorporate further measures of semantic change. SURel can be utilised to evaluate modifications and thus to optimise termhood prediction techniques regarding the sub-technical terminological meaning shift bias.

5 Extension and Discussion

We presented a gold standard for meaning shifts and how to use it for term extraction. Since our meaning shift prediction method works quite well with the however rather small dataset, we extend the target set and further compute the shifts for all nouns, verbs and adjectives in the cooking corpus with a frequency $\geq$ 50 in both SPEC and GEN. This results in shift values for 1,125 words. In the

[5] *Eiweiß, Messerspitze, Paprika, abschrecken, Strudel, Schuß, Schnee, Form, schlagen, Hamburger*

[6] Since *Messerspitze* occurred too few times in GEN, we did not compute a shift value and assumed no shift.

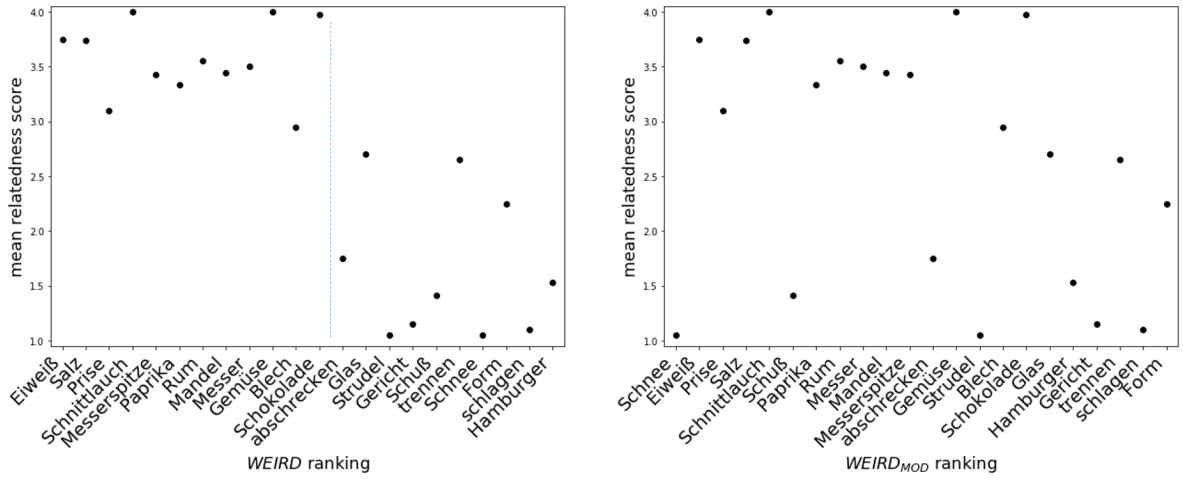

Figure 4: SURel target words ranked by WEIRD (left panel) and WEIRD$_{MOD}$ (right panel), with termhood prediction strength decreasing from left to right; the y-axes show the SURel GEN–SPEC relatedness score.

following, we use the extended dataset for remarks on challenges for term extraction.

First, our dataset contains mostly words with at least some relevance to the cooking domain. The intuition behind this is, that for clearly unterminological words (e.g. *anderes* 'different', *alternativ* 'alternative', *komplett* 'complete', *Ganze* 'whole') there should not be a meaning shift towards the domain. In practice, when applying our method, our system predicts a high degree of meaning shift for those words. Many of those words seem to be highly versatile in GEN and in SPEC. Additionally, especially problematic are words which occur without context in many cases (*Galerie* '[picture] gallery', *Inhaltsverzeichnis* 'table of contents'), or words with repeating similar context (e.g. *Wikipedia*, *Artikel* 'article', *Thema* 'topic' in the reoccurring sentence 'Wikipedia has one article to the topic ...' in the SPEC corpus). For the latter two cases, it is possible to filter the corpus beforehand, but the first case is more difficult.

We achieve some promising results with the following method: We compute a second shift value, but this time shuffle the sentences across the corpora while preserving the target word's context sentence frequencies in each corpus. By that we obtain some kind of ground truth value for the word's context variance. The assumption here is that if a word already has strongly varying contexts throughout the corpora, then the high shift across corpora is most likely a result from that. We finally substract the shuffling value from the shift value. In the resulting ranked list, this method separates the unterminologic elements to the one end and a lot of terms with meaning shift to the other end: *altbacken* 'dowdy/stale', *gedämpft* 'low voice/steamed', *Schnee*, *Fond* 'fund/stock', *Auflauf* 'crowd/casserole', *Form* 'shape/(baking) mould' together with other cooking-related words like *Spaghetti*, *Pfannkuchen* 'pancake', *Pommes* 'French fries', *Ananas* 'pineapple', where the latter words have a lower original shift value. However, other sub-technical terms like *schlagen* 'beat/whip (cream)', *abschrecken* 'discourage/chill ', *binden* 'tie/thicken (sauce)' are still among the unterminologic elements, most likely because they have rather varying contexts in GEN as well. Nevertheless, for terms with meaning shifts identified with the described method the original shift value could be used to correct a termhood measure.

6 Conclusion

We presented SURel, a German dataset for meaning shift annotations from general to domain-specific language, focusing on the language of cooking. Meaning shifts are relevant for contrastive term extraction systems, because the affected terms are typically biased towards their general-language use and, consequently, might not be recognized as terms. SURel can be used as a gold standard for predicting meaning shifts, and these predictions can be used to optimize term extraction measures. A case study incorporating a state-of-the-art diachronic semantic change measure into a simple term extraction model confirmed this potential of SURel.

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.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 451–462, Vancouver, Canada.

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.

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.

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, Malta.

Teresa Mihwa Chung and Paul Nation. 2003. Technical vocabulary in specialised texts. *Reading in a Foreign Language*, 15(2):103–116.

J Ronayne Cowan. 1974. Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly*, pages 389–399.

Damien Cram and Beatrice Daille. 2016. Terminology extraction with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, pages 13–18, Berlin, Germany.

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.

Patrick Drouin. 2004. Detection of domain specific terminology using corpora comparison. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 79–82, Lisbon, Portugal.

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.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany.

Anna Hätty, Michael Dorna, and Sabine Schulte im Walde. 2017. Evaluating the reliability and interaction of recursively used feature classes for terminology extraction. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 113–121, Valencia, Spain.

Anna Hätty and Sabine Schulte im Walde. 2018. A laypeople study on terminology identification across domains and task definitions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 321–326, New Orleans, Louisiana.

Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289.

Chunyu Kit and Xiaoyue Liu. 2008. Measuring mono-word 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.

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.

Sujatha Menon and Jayakaran Mukundan. 2010. Analysing collocational patterns of semi-technical words in science textbooks. *Pertanika Journal of Social Sciences and Humanities*, 18(2):241–258.

Antonio Moreno-Ortiz and Javier Fernández-Cruz. 2015. Identifying polarity in financial texts for sentiment analysis: A corpus-based approach. *Procedia-Social and Behavioral Sciences*, 198:330–338.

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, Hong Kong.

Ina Roesiger, 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*, pages 41–51, Osaka, Japan.

Juan C. Sager. 1990. *A Practical Course in Terminology Processing*. John Benjamins Publishing, Amsterdam.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana, USA.

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

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

Appendix

lexeme	POS	translations	MRS	freq. GEN	freq. SPEC
Strudel	NN	whirlpool, strudel (a pastry)	1.05	232	46
Schnee	NN	snow, beaten egg whites	1.05	2,228	53
schlagen	VV	beat, whip (e.g. cream)	1.10	14,693	309
Gericht	NN	court, dish	1.15	13,263	1,071
Schuß	NN	shot (e.g. gunshot, shot of milk)	1.42	2,153	117
Hamburger	NN	citizen of Hamburg, hamburger	1.53	5,558	46
abschrecken	VV	discourage, chill (with cold water)	1.75	730	170
Form	NN	shape, (baking) mould	2.25	36,639	851
trennen	VV	separate	2.65	5771	170
Glas	NN	glass (e.g. material, drinking glass, jar)	2.70	3,830	863
Blech	NN	iron plate, baking tray	2.95	409	145
Prise	NN	pinch (e.g. of humour, tobacco, salt)	3.10	370	622
Paprika	NN	bell pepper, paprika	3.33	377	453
Messerspitze	NN	point of a knive, pinch (e.g. of salt)	3.43	39	49
Mandel	NN	tonsil, almond	3.45	402	274
Messer	NN	knife	3.50	1,774	925
Rum	NN	rum	3.55	244	181
Salz	NN	salt	3.74	3,087	5,806
Eiweiß	NN	protein, egg white	3.75	1,075	3,037
Schokolade	NN	chocolate	3.98	947	251
Schnittlauch	NN	chives	4.00	156	247
Gemüse	NN	vegetable	4.00	2,696	1,224

Table 1: SURel dataset. MRS (mean relatedness score) denotes the compare rank as described in (Schlechtweg et al., 2018), where high values denote low change. Translations are illustrative for possible meaning shifts, while further senses might exist.