Originally published in: Lang, Christian/Khan, Anas Fahad/ Ostroški Anić, Ana/Spahiu, Blerina/Gracia, Jorge/McCrae, John P./Gromann, Dagmar/Heinisch, Barbara/Salgado, Ana (Eds.): Language, Data and Knowledge 2023 (LDK 2023). Proceedings of the 4th Conference on Language, Data and Knowledge – Portugal: NOVA FCSH - CLUNL, 2023. Pp. 239-244.
DOI: *https://doi.org/10.34619/srmk-injj*

Making Non-Normalized Content Retrievable – A Tagging Pipeline for a Corpus of Expert-Layperson Texts

Christian Lang and **Ngoc Duyen Tanja Tu** and **Laura Zeidler**
Leibniz Institute for the German Language
Mannheim, Germany
{lang, tu}@ids-mannheim.de
laura.zeidler@swhk.ids-mannheim.de

Abstract

Conventional terminology resources reach their limits when it comes to automatic content classification of texts in the domain of expert-layperson communication. This can be attributed to the fact that (non-normalized) language usage does not necessarily reflect the terminological elements stored in such resources. We present several strategies to extend a terminological resource with term-related elements in order to optimize automatic content classification of expert-layperson texts.

1 Introduction

One of many applications of Knowledge Organization Systems (KOS) is tagging texts to make them retrievable, cf. (Golub et al., 2019, p. 205). In our contribution, we describe the use of a KOS to process texts from the domain of expert-layperson communication – specifically, so-called language inquiries, i.e. questions that (supposed) laypeople ask linguistic experts about (German) language such as (1).

(1) Question: [...] *Muss bei... Kurs des Studienkreis_es... der Genitiv angezeigt werden, oder kann man 'Studienkreis' als undeklinierbaren Eigennamen einstufen* [...]? ([...] Does... course of the study group... need to display the genitive case, or can 'study group' be classified as an indeclinable proper name [...]?)
Answer: *Im Deutschen werden Eigennamen grundsätzlich gebeugt.* [...] *Dies gilt auch in Ihrem Beispiel.* [...] (In German, proper names are always inflected. [...] This also applies to your example [...].)

Because language inquiries serve as a valuable primary source of authentic language data for a variety of linguistic research questions, cf. (Breindl, 2016), we plan to create a monitor corpus to make them accessible to the research community. The core of this corpus is a collection of approx. 50,000 inquiries (and corresponding answers) sent by email to the language consulting service of a German publisher between 1999 and 2019.[1] The collection also contains additional metadata, such as the assignment of each question to a linguistic category (e.g. grammar, spelling, punctuation, etc.).

For optimal usability of the corpus by the research community, it is essential that researchers have access to the exact data points that are relevant to their research question. To make this possible, we identified and tagged elements in questions and answers that allow for the most precise content classification possible.

A first step in this process was terminological tagging, for which we utilized a KOS (see Section 2.1). However, as we show in Section 2.2, due to the nature of the data (expert-layperson communication), terminological tagging on its own is not sufficient. Therefore, in Section 4 we present strategies how to extend the KOS we use to meet the specific requirements of tagging texts in the domain of expert-layperson communication.

The extension of the KOS is a work in progress. Thus, we illustrate the strategies and their positive impact on the tagging process with individual example cases.

2 Tagging process

2.1 Terminological resource: WT

WT (Wissenschaftliche Terminologie)[2] is the terminological resource of the grammatical information system *grammis*.[3] It is stored and maintained in an object-relational database. The resource – an

[1] We will expand this core continuously with language inquiries received by Leibniz Institute for the German Language. In addition, we plan to extract language inquiries from other sources, including online sources, and add them to the corpus.

[2] A more exhaustive description of the resource can be found i.a. in (Suchowolec et al., 2019).

[3] https://grammis.ids-mannheim.de

onomasiologically-structured KOS that can be classified as a thesaurus according to Zeng's taxonomy of KOS (Zeng, 2008, p. 161) – contains approx. 1,900 concepts from the domain of (German) grammar. As Figure 1 shows, various attributes, such as terms or explanatory texts, can be assigned to each concept. The concepts are linked to each other using three different semantic relations: (i) as hyperonyms and hyponyms (broader term (BT) and narrower term (NT)), (ii) as holonyms and meronyms (broader term partitive (BTP) and narrower term partitive (NTP)) and (iii) as non-hierarchical relatives (related terms (RT)), cf. (ANSI/NISO Z39.19-2005 (R2010), 2005)). Currently, the resource contains 2,961 German-language and 1,874 foreign-language terms.

While terms are not restricted to nouns in principle, WT has a strong bias towards nominal terms: Approx. 90% of WT's elements are either single nouns or complex noun phrases.[4]

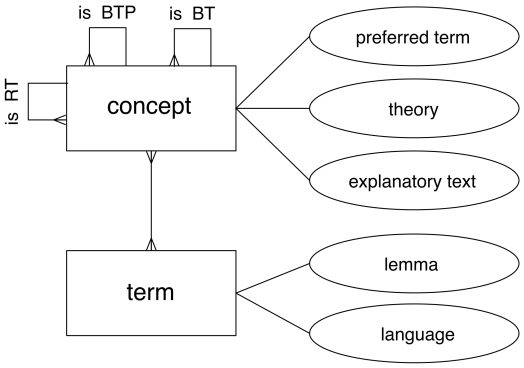

Figure 1: Data structure of WT, figure was first published in (Lang and Suchowolec, 2020, p. 31).

WT was also adapted into a SKOS vocabulary using the D2RQ platform, cf. (Suchowolec et al., 2019).

2.2 Terminological tagging

We used the terms of WT as the basis for a string-matching algorithm to tag specific keywords in our corpus. The algorithm operates as follows: First, we tokenized the data using spacy (Honnibal and Montani, 2017). Second, we applied three different lemmatizers, namely spacy, HanTa (Wartena, 2019) and GermaLemma.[5] By using multiple lemmatizers, we tried to mitigate possible weaknesses in the performance of the individual tools regarding the lemmatization of low-frequent, specialized words, namely linguistic terms. Since spacy is a look-up lemmatizer for German, it is used as a baseline, i.e., if spacy lemmatizes successfully (based on spacy's *out of vocabulary*-attribute), the lemma is adopted. If the lemmatization with spacy fails, we consult the results of the remaining two rule-based lemmatizers. If GermaLemma lemmatizes successfully, this result is adopted, otherwise we fall back on the lemmatization of HanTa.[6] If all lemmatization attempts fail, i.e. if neither lemmatizer transforms the token in any way, the token itself is adopted. Finally, we used the terms in WT – preprocessed identically to the inquiries – as the basis for string-matching to identify and tag the terms in the language inquiries.

	questions (n=800)		
	Precision	Recall	F-Measure
uni	0.74	0.898	0.811
uni & bi	0.739	0.899	0.811
uni & tri	0.74	0.9	0.812
uni, bi & tri	0.739	0.901	0.812
	answers (n=300)		
	Precision	Recall	F-Measure
uni	0.691	0.849	0.762
uni & bi	0.69	0.858	0.765
uni & tri	0.691	0.849	0.762
uni, bi & tri	0.69	0.858	0.765

Table 1: Evaluation of string-matching algorithm. The evaluation was performed on a manually annotated gold standard consisting of 800 instances of linguistic inquiries and 300 instances of corresponding answers.

We evaluated the algorithm on a subset of the corpus. To this end, two linguists created a gold standard by manually annotating elements (up to 3-grams) they deemed to be terms (for example, *Deklination* (declension), *Kleinschreibung* (lower case), etc.) in a randomly selected subset of 1,100 data points (800 questions, 300 answers). Table 1 shows the results of the evaluation, i.e. string-matching algorithm vs. manually annotated gold standard.

The evaluation reveals problems in the tagging process. On the one hand, the precision value is comparatively low. A qualitative analysis of the

[4] In their analysis of the linguistic properties of terms based on English terminological dictionaries from various technical fields, (Justeson and Katz, 1995, p. 83) found that depending on the domain, between 92% and 99% of terms are nouns or noun phrases.

[5] https://github.com/WZBSocialScienceCenter/germalemma

[6] We put GermaLemma first because this order proved to yield slightly better results in previous experiments.

elements falsely tagged as terms shows that these mainly are polysemous words that have both a technical (linguistic) and a general meaning, e.g. *Ar-gument* (argument), *Thema* (topic). On the other hand, the comparatively high recall value turns out to be deceiving on closer inspection. About 32%of all data points were not tagged at all (neither automatically nor by the human annotators) be-cause they did not contain any terms in the strict sense. Further, about 43% of data points contain either no terms or one of the two very broad terms *Satz* (sentence) and *Wort* (word) – which are un-suitable for precise content classification.[7] This result is not surprising in view of the fact that the tagged data can be attributed to the field of expert-layperson communication. That is, elements of domain-specific language do appear, but – as the following section shows – not always in the form in which they are stored in a terminological resource such as WT. It thus becomes clear that a purely terminological tagging of the data cannot guarantee optimal retrievability.

2.3 Term-related elements

We find that the data points contain elements that, while not terminology in the strict sense, may crucially contribute to the classification of the questions and answers. We refer to these elements as term-related elements. Thus, in a follow-up step, the annotators marked all term-related elements in the 1,100 data points of the gold standard.[8]

A qualitative analysis reveals broadly speaking two types of term-related elements. Type 1 elements – which account for about 53% of all elements – are adjectives (12.2%) or verbs (41.3%), of which about 90% are derivations from a nominal term (e.g. *Komparation* > *komparieren/komparierbar* (comparison > (to) compare/comparable))[9]. Type 2 elements are nouns (46% of all term-related elements), of which almost 50% are compounds or nominal phrases that have at least one term as a component (e.g. ***Genitiv**bezug* [genitive reference] or *paariges* ***Komma*** [paired comma]); another 34% of Type 2 elements are gen-eral language expressions (e.g. *Form* (form)).

If we include term-related elements, the proportion of untagged data points drops to 16% (compared to 32% when only terms are considered). In the case of data points that do not contain terms or term-related elements, linguistic examples play an important role (see Section 4.3). Although term-related elements are still insufficient to identify all questions and answers, the improvement is substantial and we believe the tagging process will benefit greatly from considering these elements.

The implementation differs depending on the type of term-related elements. While for the identification of some elements a mere adjustment in the tagging process is sufficient, for others an inclusion in WT as the KOS underlying the tagging process makes sense. For example, Type 2 compounds consisting of a term and one (or more) non-terms (e.g. *Kann**komma*** (optional comma)) can be found by partial string-matching. Including these kinds of elements in WT is not particularly useful, especially since potentially infinite compositions of terms with other words exist. However, including Type 1 derivatives in WT will not only optimize the current tagging process, but also expand the future applications of the KOS.

3 Related work

A large number of domain-specific resources of various kinds exist that can act as potential linking points for an extension of WT.

For example, LingTermNet (Neumann-Schneider and Ziem, 2020), a frame-based resource of linguistic terms containing 73 frames and 257 terms. However, the terms included are mainly from the domain of conversational analysis – an area that is less relevant to our task. Additionally, LingTermNet includes only nouns, while we want to add non-nominal elements to our resource. The latter is also true for LiDo,[10] a large relational database containing linguistic terms created by Christian Lehmann. While there are adjectives in the database, Lehmann postulates that based on conventions of scientific theory, terms should be appelatives (Lehmann, 1996, p. 4). LiDo, originally implemented in a relational database, has been converted to a Linked Data graph: LiDo RDF (Klimek et al., 2018) and is the base of OnLit, an ontology for linguistic terms

[7] The percentage of data points without terms (about 38%) and either without terms or with *Satz* (sentence) and *Wort* (word) (approx. 51%) is even higher if we consider only questions.

[8] In individual cases, it can be difficult to decide whether an element is a term or not. This classification always involves a degree of subjectivity.

[9] "Derivation from a nominal term" is not to be understood in the sense of a morphological analysis, but refers to the tendency of terms to be nouns.

[10] http://linguistik.uni-regensburg.de:8080/lido/Lido

(Klimek et al., 2017).[11]

Another approach is demonstrated by Medical WordNet, specifically for medical terms (Smith and Fellbaum, 2004), a resource that contains not only technical terms, but also medical vocabulary used by laypeople. Medical WordNet was partly built by extracting all medical terms from WordNet (Miller, 1995). WordNet is a large lexical database where among other things the semantic relation between senses of high-frequency English words is stored, either as a group of synonyms, i.e. the words refer to the same concept, or individual words.[12]

Accordingly, to extend WT, we could consider using GermaNet (Hamp and Feldweg, 1997, p. 9). While GermaNet allows different word classes to be linked (Hamp and Feldweg, 1997, p. 11), there is no noun-verb relation.[13] For example, *Deklination* (declension) and *deklinieren* ((to) decline) are not linked to each other. For that reason, GermaNet does not seem to be ideal for a systematic extension of WT. Another possible reference point is the German wiktionary.[14] We downloaded the German wikionary dump from 21-Mar-2023 00:52.[15] We extracted the titles from the wiktionary articles with a Python Package[16] and checked if WT contains the title. This is true for 1,289 titles. In some of these articles there are derivations of nominal terms, as for example in the article of *Entlehnung* (loan), where the verb *entlehnen* ((to) borrow) is listed. However, wiktionary is not domain-specific, so it is necessary to manually check whether the terms are listed in their linguistic meaning. Otherwise, it can happen that, for example, incorrect synonyms are extracted. Although some articles have the label "Linguistik" (linguistics) when a linguistic meaning is listed, not all do, such as the article for *Übersetzung* (translation).

None of the resources we considered have all the features necessary for the current task (systematic linking of nouns to other parts of speech; subject domain linguistics). Therefore, we turned to in-house resources to devise extension strategies.

[11] OnLiT offers a term-termRelation property to specify the relation of "noun Term instances and adjective and verb Term instances" (Klimek et al., 2017, p. 48-49).

[12] https://wordnet.princeton.edu/

[13] https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/lexica/germanet-1/beschreibung/relationen/

[14] https://de.wiktionary.org/wiki/Wiktionary:Hauptseite

[15] https://dumps.wikimedia.org/dewiktionary/

[16] https://pypi.org/project/wiktionary-de-parser/

4 Strategies for extending WT

With WT, we have our own comprehensive resource in which not only terms but also explanatory texts can be assigned as attributes to concepts. A total of approx. 600 terms have an explanatory text that can be used as a source for an extension. Moreover, the language inquiries themselves function as an extensive data base for finding relevant elements typically used by laypeople.

4.1 Extraction of Type 1 elements

We use (a) the terms and (b) the explanatory texts from WT to obtain Type 1 elements, i.e. derivations of nominal terms.

(a) For terms that are not linked to an explanatory text in WT, we take advantage of the fact that German is an inflectional language by applying a rule-based transformation of nominal terms in the WT into verbs and adjectives.[17] We tested these approaches with terms ending in the German noun suffix *-ung*. We chose this suffix because an analysis of the 123 verbal and adjectival term-related elements of the gold standard showed that 69% are verbs that can be nominalized by suffixation with *-ung*.[18]

(a, 1) For compounds, we automatically iterate through all terms from WT, apply a compound splitter[19] to the unigrams and filter for compounds that consist of a maximum of two elements. After that we replace *-ung* with the German verb suffix *-en* and concatenate the first constituent with the formed verb. For example, this produces *kleinschreiben* ((to) write in lowercase) for *Kleinschreibung* (lower case). Including the derived verb in the tagging process greatly increased the language inquiries found: *Kleinschreibung* yielded 1,806 language inquiries, *kleinschreiben* yielded 2,895 results (in 282 cases, both tags overlap).

(a, 2) For the remaining non-compound unigrams, we proceeded similarly, e.g. by deriving *steigern* ((to) compare) from the nominal term

[17] Rule-based approaches assume a regular derivational process, e.g. the nominalization of verbs with the suffix *-ung* or the adjectivization of verbs with the suffix *-bar*. If there is no regular relationship between noun and verb/adjective, other strategies must be applied.

[18] We also used two stemmers on the terms ending in *-ung*: while CISTEM (https://www.nltk.org/_modules/nltk/stem/cistem.html) does not correctly stem any of the terms, Snowball German Stemmer (https://www.nltk.org/_modules/nltk/stem/snowball.html) fails on 38% of the terms.

[19] https://github.com/bminixhofer/nnsplit

Steigerung (comparison). Also in this example, the integration of the verb into the tagging process leads to increased retrieval: *steigern* appears in 457 language inquiries, while *Steigerung* occurs in only 190 language inquiries (in 71 cases, both tags overlap).

(b) Next, we use the explanatory texts to find the derivations of nominal terms. We limit the search for the derivations to the explanatory texts, since in these the probability of finding true positives is very high. We perform the extraction by automatically searching for tokens in the explanatory texts that are similar to a term (with regards to spelling), have a certain suffix and belong to a certain word class (verb or adjective). For example, we search the tokenized, lemmatized and POS-tagged explanatory text of the term *Deklination* (declension) for lemmas that begin with the first three characters of the term (*dek*), end with a particular suffix, like *-ierbar* or *-ieren*, and are adjectives or verbs, depending on the suffix.[20] As a result, this produced *deklinierbar* (declinable) and *deklinieren* ((to) decline), which enabled the retrieval of 472 more language inquiries than with *Deklination* alone.

4.2 Extraction of Type 2 elements

As stated in Section 2.3, 46% of term-related elements are nouns. While approx. 50% of the 46% can be tagged by partial string-matching, a different strategy must be considered for the other half. This pertains, in particular, to general language expressions such as *Form* (form). Due to the gold standard annotations we have already a basis of term-related elements, which laypeople use instead of 'proper' linguistic terms.

These term-related elements usually occur with other words to paraphrase a linguistic term. Accordingly, we plan to perform a co-occurrence analysis on the language inquiries to analyze with which other words these elements occur frequently. We tried this approach on our gold standard data set and analyzed the co-occurrences of the token *Form*, among others, in more detail: it occurs frequently with the adjective *weiblich* (female) (164 times). We ascertained that questions containing these two words are questions about *Genus* (grammatical gender). Thus, using this methodology, we can link adjective-noun combinations to terms in WT, in this case the term-related elements *Form* and *weiblich* are linked to the concept *Genus*. This allows us to tag additional 91 language inquiries compared to tagging with *Genus* alone.

[20] We have found that the character-matching should be limited to three characters, because there are terms whose derivations could not be matched otherwise, such as *Flexion* (inflection) and *flektieren* ((to) inflect).

4.3 Extraction of examples

Terms and term-related elements do not always appear in language inquiries as stated in Section 2.3. However, in many cases an example is used in a language inquiry. Hence, on the one hand, we can extend WT with authentic examples extracted from the language inquiries, on the other hand, we can analyze the examples to identify patterns to tag them with specific terms. Therefore, language inquiries in which no terms or term-related elements are used can also be classified.

The following example of the terms *Getrenntschreibung* (separate spelling) and *Zusammenschreibung* (compound spelling) illustrates the approach: First, we clean the data by mapping all quotation marks to one quotation mark type. After that we extract the string(s) from a question that is between quotation marks, e.g. in (2), which concerns the correct spelling of "apple picking", *Apfel pflücken* (separate spelling) and *Apfelpflücken* (compound spelling) will be extracted.

(2) *Wie schreibt man "Apfel pflücken" oder "Apfelpflücken"* [...]? (How do you write "apple picking" or "applepicking" [...]?)

The strings used in questions about separate and compound spelling are identical to each other if the whitespace is removed from the separate spelling variant, as demonstrated by *Apfel pflücken* and *Apfelpflücken* in (2). Based on this pattern, we can tag 214 language inquiries from our data with the terms *Getrenntschreibung* and *Zusammenschreibung*, of which only 52 questions contained the terms or term-related elements *Getrenntschreibung*, *Zusammenschreibung*, *getrenntschreiben/getrennt schreiben* or *zusammenschreiben/zusammen schreiben*.

5 Conclusion

In our contribution, we have described the challenges that arise when using a terminological resource to tag expert-layperson texts. We have described several strategies for extending the resource. As a result, the data structure (c.f. Fig. 1) will be extended by term-related elements and language

examples (patterns).

Based on the first promising results of both the KOS extension and the adjustments in the tagging process, we suggest the following pipeline for tagging the language inquiry corpus: (1) using the entries of the extended WT to detect terms as well as term-related elements (primarily verbs and adjectives), (2) partial string-matching to identify compounds containing at least one terminological or term-related element, (3) analyzing co-occurences of term-related elements, (4) identifying typical example patterns. The next steps in optimizing the tagging process are to expand the rule-based extension beyond the cases already implemented and a systematic analysis of cases that cannot be covered by rule-based methods.

Scientific communication is assuming an increasingly more prominent role in everyday academia. This underlines the importance of creating resources and developing tools to machine process expert-layperson communication. This is why an extension of WT is a worthwhile endeavour.

References

ANSI/NISO Z39.19-2005 (R2010). 2005. ANSI/NISO Z39.19-2005 (R2010) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.

Eva Breindl. 2016. Sprachberatung im interaktiven Web. In Sven Staffeldt and Wolf Peter Klein, editors, *Die Kodifizierung der Sprache: Strukturen, Funktionen, Konsequenzen*, volume 17 of *WespA – Würzburger elektronische sprachwissenschaftliche Arbeiten*, pages 85–109. Univ. Würzburg, Institut für deutsche Philologie, Würzburg. OCLC: 959665919.

Koraljka Golub, Rudi Schmiede, and Douglas Tudhope. 2019. Recent applications of Knowledge Organization Systems: introduction to a special issue. *International Journal on Digital Libraries*, 20(3):205–207.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

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

Bettina Klimek, John McCrae, Christian Lehmann, Christian Chiarcos, and Sebastian Hellmann. 2017. Onlit: An ontology for linguistic terminology. pages 42–57.

Bettina Klimek, Robert Schädlich, Dustin Kröger, Edwin Knese, and Benedikt Elßmann. 2018. LiDo RDF: From a Relational Database to a Linked Data Graph of Linguistic Terms and Bibliographic Data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2429–2436, Miyazaki.

Christian Lang and Karolina Suchowolec. 2020. Wissensmanagement in der Praxis: Welchen Beitrag leistet deskriptive Terminologiearbeit? In Barbara Ahrens, Morven Beaton-Thome, Monika Krein-Kühle, Ralph Krüger, Lisa Link, and Ursula Wienen, editors, *Interdependenzen und Innovationen in Translation und Fachkommunikation / Interdependence and Innovation in Translation, Interpreting and Specialised Communication*, pages 17–44. Frank & Timme, Berlin.

Christian Lehmann. 1996. Linguistische Terminologie als relationales Netz. In Clemens Knobloch and Burkhard Schaeder, editors, *Nomination — fachsprachlich und gemeinsprachlich*, pages 215–267. VS Verlag für Sozialwissenschaften, Wiesbaden.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Anastasia Neumann-Schneider and Alexander Ziem. 2020. LingTermNet: Konzeption und Entwicklung eines FrameNet für linguistische Fachterminologie. In Christian Lang, Roman Schneider, Horst Schwinn, Karolina Suchowolec, and Angelika Wöllstein, editors, *Grammatik und Terminologie: Beiträge zur Ars Grammatica 2017*, number 82 in Studien zur deutschen Sprache, pages 105–128. Narr Francke Attempto, Tübingen. OCLC: on1142742225.

Barry Smith and Christiane Fellbaum. 2004. Medical WordNet: a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, pages 371–382, Geneva, Switzerland. Association for Computational Linguistics.

Karolina Suchowolec, Christian Lang, and Roman Schneider. 2019. An empirically validated, onomasiologically structured, and linguistically motivated online terminology. re-designing scientific resources on german grammar. *International Journal on Digital Libraries*, 20(3):253–268.

Christian Wartena. 2019. A probabilistic morphology model for german lemmatization. Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), pages 40–49.

Marcia Zeng. 2008. Knowledge organization systems (kos). *Knowledge Organization*, 35:160–182.