Lexical Substitution Dataset for German

Kostadin Cholakov[1], Chris Biemann[2], Judith Eckle-Kohler[3,4] and Iryna Gurevych[3,4]
(1) Humboldt University Berlin,
(2) FG Language Technology,
(3) Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Dept. of Computer Science, Technische Universität Darmstadt
(4) Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
kostadin.cholakov@anglistik.hu-berlin.de
biem@cs.tu-darmstadt.de
http://www.ukp.tu-darmstadt.de

Abstract

This article describes a lexical substitution dataset for German. The whole dataset contains 2,040 sentences from the German Wikipedia, with one target word in each sentence. There are 51 target nouns, 51 adjectives, and 51 verbs randomly selected from 3 frequency groups based on the lemma frequency list of the German WaCKy corpus. 200 sentences have been annotated by 4 professional annotators and the remaining sentences by 1 professional annotator and 5 additional annotators who have been recruited via crowdsourcing. The resulting dataset can be used to evaluate not only lexical substitution systems, but also different sense inventories and word sense disambiguation systems.

1. Introduction

The task of lexical substitution requires systems or humans to produce a substitute word for a word in a given context. For example, *lustige* ‘funny’ would be a valid substitution for the German word *humorvolle* ‘humorous’ in the following sentence:

(1) Hier werden humorvolle T-Shirts und Ansichtskarten verkauft.
‘Humorous T-shirts and postcards are sold here.’

Lexical substitution has been primarily used for evaluating the performance of Word Sense Disambiguation (WSD) systems (e.g., the English Lexical Substitution Task (McCarthy and Navigli, 2009), a part of the SemEval-2007 workshop). The main motivation behind it is to demonstrate the abilities of WSD systems on a task which has potential real-life applications in NLP. Finding alternative words which can occur in a given context would be useful to question answering, paraphrase acquisition (Dagan et al., 2006), summarisation, text simplification and lexical acquisition (McCarthy, 2002). Another motivating factor for this manner of evaluating WSD is that it is independent of any pre-defined sense inventory, cf. (Kilgarriff, 1999). Typically, WSD systems are tested on fine-grained inventories, rendering the task harder than it needs to be for many applications (Ide and Wilks, 2006). A system that performs lexical substitution, however, has a WSD component which in turn makes use of a particular sense inventory. Hence, different sense inventories of different granularities can be considered as parameters of a lexical substitution system and can be evaluated as well. Note that lexical substitution can also be performed in a cross-lingual manner (McCarthy et al., 2013).

We present a dataset designed for performing lexical substitution for German. Our main motivation for developing this dataset for German is that currently, it is not possible to test WSD systems on a sense inventory other than the fine-grained GermaNet (Hamp and Feldweg, 1997), and the recently released WebCAGe (Henrich et al., 2012) dataset annotated with GermaNet senses. Our dataset fills this gap, and its construction principle leads to a more realistic sense distribution for polysemous words. While we plan to train supervised (i.e. lexicalized (Biemann, 2012) or delexicalized (Szarvas et al., 2013)) lexical substitution systems on this data, as well as to test a resource-driven lexical substitution system on the dataset, this paper is only concerned with the description of the dataset.

The dataset includes 153 words (51 nouns, 51 adjectives, and 51 verbs) with a total of 2,040 sentences. The words have been selected based on their frequencies in large German corpora. For each part-of-speech (POS) there are 17 low-frequency words, 17 medium-frequency ones, and 17 high-frequency words. For each target noun and adjective 10 sentences have been annotated while for each verb the number of annotated sentences is 20.

The dataset is split into 2 parts. The first part contains 200 sentences and it was used for a pilot study involving professionally trained annotators. The aim of this study was to investigate the quality of the dataset, to measure initial inter-annotator agreement, and to pinpoint and fix any issues with the annotation guidelines and with the annotation interface. The study included 4 annotators who had to find substitution words for 5 nouns, 5 adjectives, and 5 verbs in a total of 200 sentences. The results of the study were very encouraging which allowed us to proceed with the annotation of the whole dataset.

The second part of the dataset includes the remaining 138 words with a total of 1,840 annotated sentences. This part is annotated by using crowdsourcing mostly. For each sentence, 5 different annotators were allowed to propose substitution words, with each annotator being allowed to process a maximum of 100 sentences. Additionally, all sentences are also annotated by a trained annotator, bringing the total number of annotators per sentence at 6. The overall quality of the annotations was very good. Nevertheless, 2 additional trained annotators, after reviewing the dataset, removed some wrong annotations and made minor corrections (e.g., correcting typos).

The remainder of the paper is organised as follows. Section 2 describes the task of lexical substitution, as it was presented to the annotators. Section 3 describes the selection process for data to be annotated. Section 4 presents the results of the pilot study. Section 5 describes the annotation of the second part of the dataset as well as the major issues with using crowdsourcing annotations. Section 6 provides details about the release format of the dataset.

2. The Task

The task of the annotators is to find a substitute for a given word in the context of a sentence. The target words include nouns, verbs, and adjectives. Unlike the SemEval-2007 English substitution task, we do not consider adverbs since there are very few "classical" adverbs in German. Instead, adjectives can almost always be used in adverbial contexts.

The annotation guidelines state three types of allowed substitutions:

1. The annotators must always try to propose a single word as a substitution
2. If it is not possible to find a single-word substitution, they are allowed to use a phrase, if the meaning of this phrase perfectly fits that of the target word
3. A slightly more general word could be used if the two options above are not applicable

Note that, for verbs, the annotators were explicitly allowed to use substitutions which alter the syntax of the sentence as long as the substitution word preserves the exact meaning of the sentence. In the following example, *unterstützt* 'to support' is a valid substitute for *hilft* 'to help' although it changes the structure of the sentence and even introduces new words:

(2) Sie ***hilft*** mir, den Bericht zu schreiben. ⟶
Sie ***unterstützt*** mich dabei, den Bericht zu schreiben.
'She helps me to write the report.' ⟶
'She supports me with the writing of the report.'

We use the freely available CrowdFlower platform[1] to specify and host the annotation interface. The platform allows for a quick setup and extensions or changes can be easily introduced. Figure 1 shows a screenshot of the interface. The annotators can provide up to 5 substitutions, but all should be fitting the context equally well. Thus, it is not required to order the substitutions based on their quality and all substitutions are viewed as a set of equally possible solutions to the task. The interface displays one sentence at a time with the target word marked in bold. The annotation guidelines are always accessible and the annotators can log in and out at any time.

In order to facilitate the calculation of inter-annotator agreement, for single-word substitutions, the annotators should fill in the base form. For nouns, this is the singular form, for adjectives – the base non-inflected form, and for verbs – the infinitive. Also, in accordance with German orthography, capitalisation is required for noun substitutions. We also require that the annotators indicate how difficult it was to find a substitution for the target word in the particular sentence they are presented with. Finally, if for some reason (e.g., lack of sufficient context, unclear usage of the word) the annotators cannot think of a good substitution, they fill in a minus '-' response in the first 'Substitution' field and mark 'impossible' for difficulty.

3. Data Selection

The full dataset comprises 2,040 sentences, with 103 target words (51 from each target POS). There are 10 sentences for each noun and adjective. Due to the higher polysemy of the verbs, we decided to extract 20 sentences per target verb. For comparison, the dataset in the SemEval-2007 English task comprises 2,010 sentences, 201 words each with 10 sentences.

We used the frequency list of German lemmas made available in the WaCKy project (Baroni et al., 2009)[2] to extract the target words. This list contains the frequencies of all lemmas in a 1.7 billion word German web corpus. For each target POS, we randomly selected words from 3 frequency groups: 17 words with frequency between 100 and 500, 17 with frequency between 501 and 5,000, and 17 words with more that 5,000 occurrences in the web corpus. This way, we can examine the impact frequency has on the performance of the system tested on the dataset.

Then, for each target word, we randomly extracted sentences containing this word from the German Wikipedia.[3] We used Wikipedia instead of the German corpus in WaCKy due to possible copyright issues which might arise when we make our data publicly available. Sentences with less than 6 words were excluded. The selected sentences were manually checked for well-formedness and any other problems. A random selection ensures that the sense distributions in our dataset match the sense distributions in the corpus.

4. Part I: Pilot Study

4.1. Study Setup

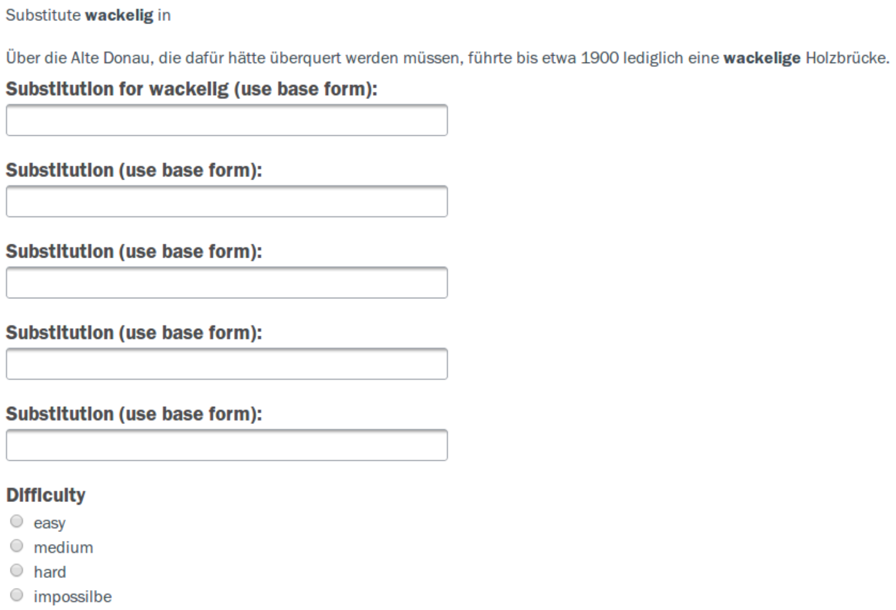
Before releasing the whole dataset, we performed a pilot study involving 4 human subjects (2 male and 2 female) and 200 sentences for 5 words of each target POS. The 15 words were randomly selected.

All subjects are native speakers of German from Germany. Further, all have educational background in linguistics. We also involved non-linguists in the annotation of the whole dataset but it was important for us to have trained annotators in the pilot study in order to pinpoint possible problems. The annotators were allowed to use dictionaries and no time limit was set for the annotation. However, we did ask each annotator to provide us with the total amount of time he or she spent on performing the task.

[1] http://www.crowdflower.com

[2] http://wacky.sslmit.unibo.it/doku.php?id=frequency_lists

[3] http://www.de.wikipedia.org



Substitute **wackelig** in

Über die Alte Donau, die dafür hätte überquert werden müssen, führte bis etwa 1900 lediglich eine **wackelige** Holzbrücke.

Substitution for wackelig (use base form):

Substitution (use base form):

Substitution (use base form):

Substitution (use base form):

Substitution (use base form):

Difficulty

○ easy
○ medium
○ hard
○ impossilbe

Figure 1: Annotation interface for the German lexical substitution task.

Last, the sentences were randomly presented to the annotators, i.e. the sentences for a given target word were not presented sequentially as a group. This was done in order to prevent the annotators from ignoring the context and proposing the same substitutions they proposed for the first sentence they saw for the target word.

4.2. Results and Analysis

The annotators reported that the guidelines were clear and the interface is easy to use. Task completion times ranged from 4 to 10 hours for 200 sentences. Table 1 shows the difficulty judgements given by all 4 annotators. The majority of the judgements are 'easy' or 'medium'. It is important to note that for all sentences where an annotator could not find a good substitution, at least 2 other subjects were able to do so. Therefore, we have not excluded any sentences from the dataset. Although the numbers in Table 1 give some indication about the difficulty of the various sentences, they should not be taken at face value. The annotators admitted that they spent more time on the first one or two sentences they saw for a given target word and they tended to mark those sentences with a higher degree of difficulty. Once the annotators had thought of substitutions, it was generally easier for them to process the remaining sentences they see for this word (despite those not being presented sequentially). Therefore, they would mark a sentence as easier although, in reality, this sentence might be harder to process than the first sentence for this target word.

Next, we calculated inter-annotator agreement (IAA) in the same manner it had been done for the Semeval-2007 English lexical substitution task. Since we have sets of substitutions for each annotator and for each sentence in our study, we calculated pairwise agreement between each pair of sets ($p_1, p_2 \in P$) as:

$$\frac{\sum_{p_1,p_2 \in P} \frac{p_1 \cap p_2}{p_1 \cup p_2}}{|P|} \quad (3)$$

Difficulty	Number of judgements
easy	240
medium	383
hard	121
impossible	51
total	800

Table 1: Difficulty judgements for the 200 sentences in the pilot study.

where P is the set of all possible pairings. In our study $|\ P\ |$ = 1200, 6 possible pairs of annotators × 200 sentences.

The calculated pairwise agreements are given in Table 2. Pairwise IAA for all words was 16.95%, with nouns having the highest agreement score. Agreement is low compared to a task with a fixed inventory. This is due to the fact that there is no clear right or wrong answer for many items but rather several possibilities, each of which has different "popularity" among the annotators. Note also that the lowest agreement is achieved for adjectives, presumably because there is typically a larger variety of potential substitutions for adjectives compared to the other target POS. Our results reflect the findings in the English substitution task where the nouns also had the highest agreement score, followed by the verbs and by the adjectives. The difference in the agreement scores for verbs and adjectives there was minimal, as it is also the case in our study.

We noticed that one of the annotators was in very low agreement with the rest of the group. Unlike the other 3 subjects, the annotator preferred phrase substitutions for many of the sentences in which the target word was a verb as well as for quite some number of sentences in which an adjective had to be substituted. If we leave this annotator out, IAA increases to 22%, which is only slightly lower than the Semeval 2007 lexical substitution task IAA on nouns verbs

POS	Number of Sentences	IAA (%)
Nouns	50	30.14
Adjectives	50	11.75
Verbs	100	12.97
All	200	16.95

Table 2: Inter-annotator agreement by POS for the 15 words in the pilot study.

and adjectives across 5 annotators, reported as 26% in (McCarthy and Navigli, 2009).

5. Part II: Crowdsource Annotation

5.1. Setup

The pilot study indicates that our annotation guidelines and our procedure for constructing the lexical substitution dataset is valid and feasible. Based on this, we decided to proceed with the annotation of the remaining 1,840 sentences by using just one trained annotator. Further, 5 additional annotators per sentence are recruited via crowdsourcing, thus bringing the total number of annotators per sentence to 6. Our main motivation for using crowdsourcing is to obtain judgements from average, non-professional people. Such people are potential users of real-life NLP applications which include lexical substitution components. Nevertheless, since quality of crowdsource data is often questionable, we decided to include a professional annotator as well.

All annotations were performed within the Crowdflower platform. The trained annotator is a male German native speaker from Germany, with a strong linguistic background. He performed the annotation in approximately 40 hours.

Next, the following requirements were set for the crowdsourcing annotators. First, only IP addresses from Germany, Austria, and Switzerland were allowed. Second, each annotator was allowed to annotate a maximum of 100 sentences, with a time limit of 5 minutes per sentence. Furthermore, the whole task description was given in German. Annotators were paid 0.05$ per sentence.

The data was split randomly into 5 portions which were then consecutively presented for annotation. The annotation process for each portion was carefully observed in order to prevent malicious annotators from misusing the interface and entering "garbage" annotations . The annotation of all 1,840 sentences took a total of 5 days. Usually, when a portion was uploaded to Crowdflower and made available for annotation, it was processed in 16 to 20 hours.

5.2. Analysis and Post-processing of Crowdsourced Data

The crowdsourcing annotators were given the same instructions as the professional ones. However, the instructions were written in such a way, so they are clear to non-professionals as well. The terms were kept to the necessary minimum and explained by clear examples. The overall quality of the annotations obtained via crowdsourcing was very good.

However, in order to assure the high quality of the data, 2 professional annotators performed post-processing during which minor changes and corrections were made. The most prominent issue was typos which we have corrected to the best of our efforts. Another prominent issue was the apparent use of online thesaurus platforms and tools for the automated generation of synonyms by some annotators. The generated lists were then used to input substitution words, irrelevant of the context in which the target word occurred. All annotators were allowed to consult dictionaries and other assisting resources but such a blind paste of whole lists of synonyms leads to more noise in the data. In the cases in which the substitutions provided were the obvious result of pasting from resources, we removed only substitutions which were completely inadequate. However, the remaining substitutions proposed were left intact.

Another issue with the crowdsourced data is that, despite clear instructions accompanied by examples, annotators have sometimes proposed substitution words the POS of which is different from the POS of the target word. Probably, not all annotator have fully grasped the notion of POS or understood the annotation instructions. This is a common problem in crowdsourcing. We have removed such substitutions from the annotations. Finally, for some nouns (e.g., *Terroristin* (Terrorist.FEM)), annotators have entered substitution words of the opposite gender. Such words were left unchanged since they reflect the different gender awareness which a male and a female annotator have and thus, they are also valid substitutions.

Despite the removal of some annotations, there remains a sufficient amount of annotations in our dataset. Each target word instance has at least 3 substitution words provided. Since each annotator can provide up to 5 substitution words or phrases, theoretically, there can be as many as 30 different substitutions suggested per instance. Although interannotator agreement cannot be measured due to the many different participating annotators, we observed, similar to the situation in the pilot study, that the substitutions for many instances were very sparse. This demonstrates once more the difficulty of the lexical substitution task.

In the next section, we describe the release format of the dataset.

6. Data Format and Release

The lexical substitution dataset contains two types of files: XML files consisting of the sentences containing the target words (ending in *.xml*) and gold files providing the substitutes suggested by the annotators along with their judgements (ending in *.gold*). The files are encoded in UTF-8 and *.xml* files contain a well-formed XML format similar to the English lexical substitution task which can be easily processed automatically. In order to illustrate the format, we show a small excerpt of an *.xml* file in Figure 2.

All instances of a given target word are grouped under a <*lexelt*> element. This element has a special attribute *item* the value of which encodes the target word together with its POS. Each instance of a given target word is encoded in an <*instance*> element with a special attribute *id* providing a unique identifier for each instance. The sentence containing the target word is within a <*context*> element, with the

```
<lexelt item="Abbuchung.n">
        ...
        <instance id="Abbuchung_2">
                <context>
                        Wenn die kartenausgebende Bank online autorisiert,
                        werden Abhebungen bei Bankomaten unmittelbar am Konto
                        verbucht, bei Fallback-Autorisierung erfolgt die
                        <head>Abbuchung</head> in der Regel binnen 2 Tagen.
                </context>
        </instance>
        <instance id="Abbuchung_3">
                <context>
                        Finanziell verzichtet Eckankar auf jede Verpflichtung
                        oder <head>Abbuchung</head> fr die Mitglieder zur
                        Zahlung von Geldern.
                </context>
        </instance>
        ...
</lexelt>
```

Figure 2: Data format for the German lexical substitution dataset.

instance of the target word marked clearly in a <*head*> element.

The substitution words or phrases suggested by the annotators for each instance are provided in *.gold* files. The format of these files also follows that of the English lexical substitution task. Figure 3 illustrates the format by showing the annotations for the two example sentences in Figure 2.

The *.gold* files contain 3 tab-separated columns. The first column contains the target word together with its POS. The second column specifies the identifier for the particular instance for which substitutions were provided. The third column lists the substitutions themselves sorted by the total number of times each substitution was proposed by the annotators. Note that the sequence in which the annotators entered the various substitutions for a given instance does not play any role in the sorting.

As we plan to use the lexical substitution dataset in a shared task, we split the dataset into a training and a test portion. The test data include all sentences containing instances of 25 nouns, 25 adjectives, and 25 verbs which we randomly selected out of the 153 target words in the dataset. Thus, the test data portion comprises a total of 1,000 sentences which is nearly half of the dataset. For now, we will hold back these 1,000 sentences and release only the remaining 1,040 sentences that form the training data.[4]

After the completion of the shared task, the whole dataset will be available as well as 4 additional files. The first pair, *pilot.xml* and *pilot.gold*, contains the data and the annotations from the pilot study. The other pair of files, *crowd.xml* and *crowd.gold* includes the data and annotations obtained by crowdsourcing. The users of the dataset thus have the opportunity to choose to rely on professional or crowdsourced data.

7. Conclusion and Outlook

We presented a lexical substitution dataset for German. The dataset contains 153 nouns, adjectives, and verbs with a total of 2,040 sentences containing instances of those words. First, we performed a pilot study with 15 words (200 sentences) involving professional annotators in order to fine-tune the conditions we set for the task. Reported interannotator agreement was consistent with the findings for a comparable English lexical substitution task which demonstrates the difficulty of the task due to lexical variability. Based on the results of the study, we proceeded with annotating the remaining 1,840 sentences by employing a single professional annotator and 5 additional annotators recruited via crowdsourcing. Some post-processing of the crowdsourced data was performed to remove or correct some wrong substitutions. As a result, each instance of a target word is provided with at least 3 substitution words or phrases. The dataset will be publicly available under a permissive CC license. For now, we have released 1,040 sentences which are to serve as training data in a shared task.

In future research, we plan to use this dataset for testing different sense inventories and WSD systems. We will also develop and propose various evaluation metrics and routines. In conclusion, we believe that our resource will be a significant contribution to NLP for German, and can serve as a dataset for a shared task.

8. References

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

Biemann, C. (2012). Creating a System for Lexical Substitutions from Scratch using Crowdsourcing. *Language Resources and Evaluation: Special Issue on Collaboratively Constructed Language Resources*, 46(2).

[4] The training portion of the dataset can be found at https://www.ukp.tu-darmstadt.de/data/lexical-substitution/lexical-substitution-dataset-german/.

```
Abbuchung.n Abbuchung_2 :: Belastung 2; Zahlung 2; Anzug 1; Auszahlung 1;
                           Liquidierung 1; Kontobelastung 1; Abzug 1; Abhebung 1;
Abbuchung.n Abbuchung_3 :: Abgang 1; Einstellung 1; Entnahme 1; Abhebung 1;
                           Einzug 1; Abschreibung 1; Abschpfungsauftrag 1;
                           Zahlungsausgang 1; Kontobelastung 1; Ende 1;
                           Beendigung 1;
```

Figure 3: Format of the substitution words or phrases proposed by the annotators.

Dagan, I., Glickman, O., Gliozzo, A., Marmorshtein, E., and Strapparava, C. (2006). Direct word sense matching for lexical substitution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 449–456.

Hamp, B. and Feldweg, H. (1997). GermaNet – a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain.

Henrich, V., Hinrichs, E., and Vodolazova, T. (2012). WebCAGe: a web-harvested corpus annotated with GermaNet senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 387–396, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ide, N. and Wilks, Y. (2006). Making sense about sense. In *Word Sense Disambiguation*, pages 47–73. Springer.

Kilgarriff, A. (1999). I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

McCarthy, D. and Navigli, R. (2009). The English lexical substitution task. *Language resources and evaluation*, 43(2):139–159.

McCarthy, D., Sinha, R., and Mihalcea, R. (2013). The cross-lingual lexical substitution task. *Language Resources and Evaluation*, 47(3):607–638.

McCarthy, D. (2002). Lexical substitution as a task for WSD evaluation. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions-Volume 8*, pages 109–115.

Szarvas, G., Biemann, C., and Gurevych, I. (2013). Supervised all-words lexical substitution using delexicalized features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 1131–1141, Stroudsburg, PA, USA, June. Association for Computational Linguistics.