Sense-annotating a Lexical Substitution Data Set with Ubyline

Tristan Miller, Mohamed Khemakhem, Richard Eckart de Castilho, and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
https://www.ukp.tu-darmstadt.de/

Abstract

We describe the construction of GLASS, a newly sense-annotated version of the German lexical substitution data set used at the GERMEVAL 2015: LEXSUB shared task. Using the two annotation layers, we conduct the first known empirical study of the relationship between manually applied word senses and lexical substitutions. We find that synonymy and hypernymy/hyponymy are the only semantic relations directly linking targets to their substitutes, and that substitutes in the target's hypernymy/hyponymy taxonomy closely align with the synonyms of a single GermaNet synset. Despite this, these substitutes account for a minority of those provided by the annotators. The results of our analysis accord with those of a previous study on English-language data (albeit with automatically induced word senses), leading us to suspect that the sense–substitution relations we discovered may be of a universal nature. We also tentatively conclude that relatively cheap lexical substitution annotations can be used as a knowledge source for automatic WSD. Also introduced in this paper is Ubyline, the web application used to produce the sense annotations. Ubyline presents an intuitive user interface optimized for annotating lexical sample data, and is readily adaptable to sense inventories other than GermaNet.

Keywords: lexical substitutions, word senses, word sense annotation

1 Introduction

Word sense disambiguation (WSD)—the task of determining a word's meaning in context—is one of the oldest open research problems in computational linguistics. WSD systems are commonly evaluated by having humans mark up the words in a text with their contextually appropriate meanings, as enumerated by a dictionary or other lexical-semantic resource, and then comparing these annotations against those supplied by the systems. Such "*in vitro*" evaluations are popular because they are straightforward to conduct, though they have the disadvantage of requiring considerable effort to produce the manually annotated gold standard data. Sense-annotated data sets remain rare, particularly for languages other than English.

A more recent and increasingly popular evaluation method, which has the advantage of not requiring all human and machine annotators to use the same sense inventory, is lexical substitution. Here the lexical annotations which are applied and compared are not sense labels, but rather lists of plausible synonyms. It has been argued that, since the identification and ranking of these substitutions depends on a proper understanding of the word's meaning in context, accuracy in this "*in vivo*" task is an indirect measure of WSD performance (McCarthy, 2002).

The contributions of the present work are twofold. First, to ease the considerable technical and ergonomic burdens of creating sense-annotated data, we provide Ubyline,[1] an Apache-licensed, web-based sense annotation tool whose user interface is optimized for lexical sample data. (*Lexical sample* data sets feature documents or short texts in which annotations are applied to all instances of a fixed set of lemmas.) Ubyline supports a wide range of sense inventories in several languages, including WordNet and GermaNet. It is, to our knowledge, the only published GermaNet-compatible sense annotation tool.

Our second contribution is GLASS (German Lexemes Annotated with Senses and Substitutions),[2] a new German-language sense-annotated data set. GLASS fills a gap in German-language resources by providing a lexical sample data set which (i) features high-quality, manually applied sense annotations, (ii) is well balanced with respect to the target words' frequency and part of speech, (iii) is of sufficient size to be useful for machine learning, and (iv) is distributed under a free content licence. Because GLASS extends an existing lexical substitution data set, it allows for *in vitro* and *in vivo* evaluations of WSD systems to be carried out on the same data. Moreover, GLASS is the first resource in any language to permit an empirical study of the relationship between manually annotated word senses and lexical substitutes.

2 Previous Work

2.1 Data Sets

There exist a handful of previously published German-language sense-annotated data sets, all of which are of the lexical sample variety. The properties of these data sets, as well as those of our own GLASS, are summarized in Table 1.

The earliest of these resources, the MuchMore evaluation set (Raileanu et al., 2002), has GermaNet annotations for 2421 occurrences of 25 nouns in a corpus of medical abstracts. Raw interannotator agreement was high (0.841). Though the authors have made the data available for download, there is no explicit statement of the terms of use, so it cannot be presumed to be freely licensed.

The DeWSD resource (Broscheit et al., 2010) has manual sense annotations for 1154 occurrences of 40 lemmas (6 adjectives, 18 nouns, 16 verbs) in the deWaC corpus (Baroni et al., 2009). While the lemmas were translated from an English WSD data set, some attempt was made to yield a

[1] https://github.com/UKPLab/lrec2016-ubyline

[2] Available at https://www.ukp.tu-darmstadt.de/data/ under the CC BY-SA 3.0 licence.

	WikiCAGe	WebCAGe	MuchMore	DeWSD	TüBa-D/Z	GLASS
Lemmas:						
adj.	0	211	0	6	0	51
nouns	1 030	1 499	25	18	30	51
verbs	0	897	0	16	79	51
total	1 030	2 607	25	40	109	153
Tokens	24 344	10 750	2 421	1 154	17 910	2 038
GermaNet ver.	6.0	7.0–9.0	1.0(?)	5.1, 9.0	8.0	9.0
Domain	open	open	medical	open	open	open
Annotations	semi-automatic	semi-automatic	manual	manual	manual	manual
Licence	unpublished	CC BY-SA/ proprietary	proprietary	proprietary	proprietary	CC BY-SA

Table 1: Comparison of sense-tagged corpora for German

good distribution across parts of speech, polysemy, and word frequency. No information is provided on the manual annotation process, including interannotator agreement. Though the original DeWSD data set was annotated with senses from GermaNet 5.1, Henrich (2015) later updated these to GermaNet 9.0. As with the MuchMore data set, DeWSD is available for download but with no specified licence.

WebCAGe (Henrich et al., 2012b; Henrich, 2015) is a collection of 10 750 occurrences of 2607 lemmas which have been semi-automatically tagged with senses from GermaNet 7.0 through 9.0. The source contexts are all web-harvested, and include a mix of free and proprietary content. The portion of the data set derived from free sources (9376 tagged word tokens) is distributed under the terms of the CC BY-SA licence; as the full data set includes proprietary content, it is not publically available. A parallel project, WikiCAGe (Henrich et al., 2012a), applied GermaNet 6.0 sense annotations semi-automatically to a Wikipedia corpus containing 24 334 occurrences of 1030 lemmas. While it was intended to be released under a free content licence, it was never published.[3]

Recent versions of the TüBa-D/Z treebank (Henrich and Hinrichs, 2013; Henrich and Hinrichs, 2014) include manually applied GermaNet 8.0 annotations for 17 910 occurrences of 109 lemmas (30 nouns and 79 verbs). Lemmas were selected to ensure a good balance of word frequencies, number of distinct senses, and (for verbs) valence frames. Interannotator agreement was generally good (mean Dice coefficient of 0.964 for nouns and 0.937 for verbs). While the data is available for non-profit academic use, it is not released under a free content licence.

With respect to German-language lexical substitution data sets, the only one of which we are aware is that of Cholakov et al. (2014). As the present work greatly builds upon its existing content, we reserve our description of it for §3.

2.2 Annotation Tools

Manual linguistic annotation, and sense annotation in particular, is known to be a particularly arduous and expensive task (Mihalcea and Chklovski, 2003). The process can be facilitated through the use of dedicated annotation support software. Several tools have been developed for applying WordNet (Fellbaum, 1998) senses to English text. SATANiC (Passonneau et al., 2009), for example, was used to build the MASC corpus; Punnotator (Miller and Turković, 2016) was created specifically to support WordNet sense annotation of English puns. The only multilingual sense annotation tool we are aware of, IMI (Bond et al., 2015), was used to annotate the NTU-Multilingual Corpus (Tan and Bond, 2011) with senses from the Open Multilingual Wordnet (OMWN) (Bond and Foster, 2013). Although WordNet, OMWN, and GermaNet share a similar structure, the aforementioned tools do not support GermaNet (nor any other German sense inventory).

To our knowledge, only two sense annotation tools work with GermaNet. The first of these is KiC, which was used to produce the MuchMore data set. It does not appear to have been publically released, though a brief description appears in Raileanu et al. (2002). KiC displays sentences for a given target word in KWIC (key word in context) format alongside a list of candidate senses from GermaNet. Annotators select the appropriate senses for each occurrence of the target word; they also have the option of marking an occurrence as “unspecified” if GermaNet does not contain its sense. To help distinguish between problematic senses, KiC can show their corresponding hypernym–hyponym hierarchies.

The second GermaNet-capable annotation tool is an unnamed browser-based interface used to extend the TüBa-D/Z treebank. As with KiC, it has not been published, though it is described by Henrich (2015). This tool displays target word occurrences in their sentential context, though separately rather than in KWIC format. Below each context is a list of candidate senses, identified only by their GermaNet numeric IDs and a brief description. If this information is not sufficient to discriminate between the senses, annotators must use a separate GermaNet exploration tool such as GernEdiT (Henrich and Hinrichs, 2010). Users select senses by clicking on them; for problematic cases, there is a text field to type natural-language comments.

3 Resources

GLASS applies sense annotations to the lexical substitution data set previously described by Cholakov et al. (2014) and later released in full under the Creative Commons

[3] Personal communication with V. Henrich, 7 September 2015.

Attribution-ShareAlike 3.0 Unported licence (Miller et al., 2015). Our decision to use this data set was motivated by its free licensing, and by the fact that having both sense and lexical substitution annotations will allow for intrinsic and extrinsic evaluations of WSD systems to be carried out on the same data. Moreover, the double annotations provide a rich resource for investigating the relationship between word senses and lexical substitutions. (The only previous study on this topic, Kremer et al. (2014), uses automatically induced rather than manually applied word sense annotations.)

The Cholakov et al. (2014) data is provided as XML and delimited text files, and consists of 2040 context sentences from the German edition of Wikipedia, each containing one annotated target word. There are 153 unique target words, equally distributed across parts of speech (nouns, verbs, and adjectives) and three frequency bands as measured by word frequency counts in the German deWaC corpus (Baroni et al., 2009). The data set's creators did not control for polysemy or synonymy as they did not wish to introduce a bias towards any one sense inventory. There are ten context sentences for each noun and adjective and twenty for each verb. A list of contextually appropriate substitutions is provided for each target word; 200 targets were annotated by four professional human annotators, and the remaining 1840 by one professional annotator and five additional annotators recruited via crowdsourcing. The data set has already seen use in an organized lexical substitution evaluation exercise (Miller et al., 2015).

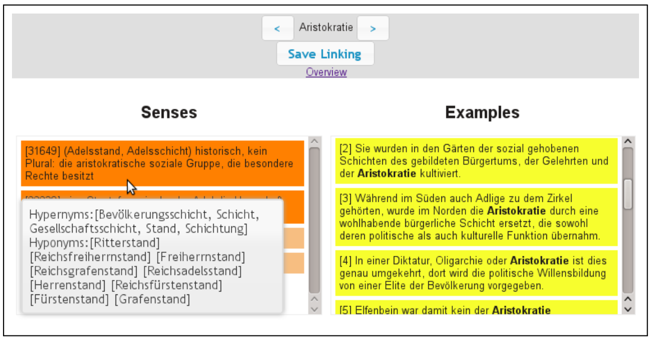
Our review of the original lexical substitution data set revealed that two of the lemmas had duplicate context sentences. We removed these, lowering the total number of contexts in the data set to 2038. We also corrected a number of inconsistencies in the segmentation of words.

In line with the sense-annotated corpora discussed previously, we chose GermaNet as our sense inventory. GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) is a lexical-semantic network that relates German-language nouns, verbs, and adjectives. Like its English analogue, WordNet, GermaNet represents semantic concepts as *synsets* which are interlinked through labelled semantic relations. We used version 9.0 of the resource, which contains 93 246 synsets covering 121 810 lexical units.

4 Ubyline

As discussed in §2.2, most existing sense annotation tools do not support using GermaNet as the sense inventory, and those that do are unpublished. We therefore developed Ubyline, a web-based sense annotation tool. Our tool is a Java-based web application; it uses CQP (Evert and Hardie, 2011) for querying, and MySQL for managing the sense inventory and storing the user annotations.

Data Preparation. Ubyline is able to import a sense inventory from any of the lexical resources supported by UBY (Gurevych et al., 2012), including GermaNet. Since the GermaNet licence does not allow redistribution, we build a UBY lexicon from GermaNet using *ubycreate*.[4] Ubyline employs DKPro Core (Eckart de Castilho and Gurevych, 2014) to read the corpus data, mark lemmas to be annotated, segment the texts, and create the CQP indexes.



[31649] (Adelsstand, Adelsschicht) historisch, kein Plural: die aristokratische soziale Gruppe, die besondere Rechte besitzt

Hypernyms:[Bevölkerungsschicht, Schicht, Gesellschaftsschicht, Stand, Schichtung]
Hyponyms:[Ritterstand] [Reichsfreiherrnstand] [Freiherrnstand] [Reichsgrafenstand] [Reichsadelsstand] [Herrenstand] [Reichsfürstenstand] [Fürstenstand] [Grafenstand]

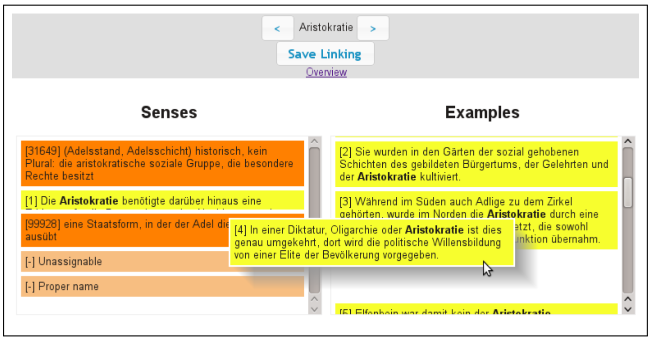
[2] Sie wurden in den Gärten der sozial gehobenen Schichten des gebildeten Bürgertums, der Gelehrten und der **Aristokratie** kultiviert.

[3] Während im Süden auch Adlige zu dem Zirkel gehörten, wurde im Norden die **Aristokratie** durch eine wohlhabende bürgerliche Schicht ersetzt, die sowohl deren politische als auch kulturelle Funktion übernahm.

[4] In einer Diktatur, Oligarchie oder **Aristokratie** ist dies genau umgekehrt, dort wird die politische Willensbildung von einer Elite der Bevölkerung vorgegeben.

[5] Elfenbein war damit kein der **Aristokratie**

Figure 1: Hovering over a sense in Ubyline



[31649] (Adelsstand, Adelsschicht) historisch, kein Plural: die aristokratische soziale Gruppe, die besondere Rechte besitzt

[1] Die **Aristokratie** benötigte darüber hinaus eine

[99928] eine Staatsform, in der der Adel die
ausübt

[-] Unassignable

[-] Proper name

[2] Sie wurden in den Gärten der sozial gehobenen Schichten des gebildeten Bürgertums, der Gelehrten und der **Aristokratie** kultiviert.

[3] Während im Süden auch Adlige zu dem Zirkel gehörten, wurde im Norden die **Aristokratie** durch eine
etzt, die sowohl
unktion übernahm.

[4] In einer Diktatur, Oligarchie oder **Aristokratie** ist dies genau umgekehrt, dort wird die politische Willensbildung von einer Elite der Bevölkerung vorgegeben.

[5] Elfenbein war damit kein der **Aristokratie**

Figure 2: Drag-and-drop for sense linking in Ubyline

User Interface. Upon log-in, Ubyline presents an overview of lemmas and occurrences requiring annotation. Clicking on a lemma brings up the annotation interface showing candidate senses and sentences containing the target lemma ("examples") in side-by-side lists. Senses are represented by their GermaNet ID, synonyms, and gloss. Where the distinction between senses remains unclear (which is often the case, as GermaNet does not provide glosses for every sense), a tooltip informs the user about hypernym and hyponym synsets (see Figure 1). Following the practice of past sense-annotated data sets, Ubyline also provides two special senses, "unassignable" and "proper name".

To apply a sense annotation, the user drags an example from the right-hand list and drops it at position under the appropriate sense on the left (see Figure 2). This action does not remove the example from the right-hand list, allowing one example to be assigned to multiple senses. Existing annotations can be modified or removed by dragging them to a new position in the sense list or back to the example list, respectively. Once all examples have been assigned to at least one sense, the user can proceed to the next lemma, or return to the lemma overview. The disambiguation performed by each annotator can be directly applied to the original data set file and downloaded by a simple click on an "Export" button in the lemma overview page.

Comparison. Both Ubyline and KiC have major advantages over the TüBa-D/Z annotation tool: both present all senses and instances of a given lemma simultaneously (allowing the annotator to better survey and distinguish between their meanings), and both present detailed sense infor-

[4] https://github.com/dkpro/dkpro-uby/tree/master/de.tudarmstadt.ukp.uby.ubycreate-gpl

mation from GermaNet rather than requiring tedious switching between applications. But unlike KiC, which is a standalone binary application, the TüBa-D/Z tool and Ubyline are client–server applications which annotators can easily access from any web browser. Ubyline also has several features not present in either of the other two tools: it is freely licensed, it has a highly intuitive user interface, and it produces fine-grained logging of all user actions.

The reliance on UBY and DKPro Core provides Ubyline with access to various sense inventories, corpus formats, and preprocessing steps. For example, UBY integrates eleven resources in various languages. However, adapting Ubyline to new corpora presently requires changing the import code, in particular to mark the target words in each sentence.

5 Annotation Process

We trained and engaged three human judges—all native German-speaking graduate students in computational linguistics—to produce our manually annotated data set. Two judges independently sense-annotated all 2038 instances in our data set, and the third served as an adjudicator who went through all occurrences where the two annotation sets differed and resolved the disagreements.

The two annotators were trained in the use of Ubyline and given oral and written annotation guidelines. We configured Ubyline such that, in addition to the senses from GermaNet, annotators had the option of applying two special senses: "proper name" (P) for senses not in GermaNet because they refer to a proper name, and "unassignable" (U) for senses not in GermaNet for any other reason. The annotators were free to consult outside sources, such as dictionaries, to help them understand the contexts, but they were not permitted to discuss cases with each other.

After annotating twenty lemmas each, the guidelines were revised to account for some anomalies in the data and in GermaNet itself. For instance, the annotators had discovered that for at least one lemma, *Korrektur*, the definition given by GermaNet was more specific than the hyponyms it lists. On further investigation, this appeared to be an error introduced by a recent project to supplement GermaNet's definitions with those semi-automatically extracted from Wiktionary (Henrich et al., 2014). We therefore instructed the annotators to resolve any apparent conflicts between GermaNet's sense definitions and hypernym–hyponym taxonomy in favour of the latter.

For the adjudication phase, the adjudicator was provided with the original annotation guidelines as well as a set of adjudication guidelines. The latter basically instructed the adjudicator, for each instance on which the annotators disagreed, to accept one or the other set of annotations, or the union of the two. A custom browser-based interface was provided to effect the adjudications; this presented similar information as Ubyline, plus which senses (or sets thereof) were selected by the two annotators.

6 Analysis

6.1 Interannotator Agreement

Following Raileanu et al. (2002) and Henrich and Hinrichs (2013), we calculate interannotator agreement (IAA) using both raw percentage agreement and the Dice coefficient. Our mean raw agreement is high (0.861, 0.865, and 0.815 for adjectives, nouns, and verbs, respectively), and seemingly better than that reported for MuchMore (0.841 for nouns). The Dice coefficient, which awards credit for partial matches, gives us IAA scores of 0.873, 0.896, and 0.835 for adjectives, nouns, and verbs, respectively. These results are somewhat lower than those reported for TüBa-D/Z (0.964 for nouns and 0.937 for verbs), though it should be noted that unlike our annotators, theirs did not have the option of marking target words as unassignable or proper names. Furthermore, because these measures of IAA do not account for the sense distributions within and across data sets, they may not meaningfully reflect the relative reliabilities of the data sets.

Both Raileanu et al. (2002) and Henrich (2015) make further computations of IAA per lemma using Cohen's κ, a chance-correcting measure of IAA. However, this metric is unable to cope with instances that receive multiple sense annotations (as happened in about 4.1% of our cases, as well as 3.3% and 0.4% of the MuchMore and TüBa-D/Z instances, respectively). Furthermore, neither Cohen's κ, nor other IAA measures which do work with multiple labels (such as Krippendorff's α), return meaningful results when all annotators apply the same sense annotation to all occurrences of a given lemma. This situation arises relatively often for our lemmas, which have lower average polysemy and much lower occurrence counts than those of TüBa-D/Z and MuchMore.

Raileanu et al. (2002) and Henrich and Hinrichs (2013) skirt both problems by simply excluding the affected instances from their κ calculations. With this expediency, the MuchMore lemmas yield κ scores ranging from 0.33 to 1.00, and the TüBa-D/Z ones from −0.00 to 1.00. When we do likewise, we observe a much wider range of κ scores, from −0.43 to 1.00. Since there were no obvious patterns of disagreement in the early phase of the annotation process, we suspect that this is a result of differences in our data set rather than an indicator of low quality. That is, the lemmas with systematic disagreement are indeed an artifact of their low polysemy and lower applicable occurrence counts. As further evidence of this, we observe a moderate negative correlation between lemma polysemy and (Dice) agreement for adjectives and nouns, with Pearson's $r = -0.302$ and -0.333, respectively. There is, however, no appreciable correlation for verbs ($r = -0.076$).

In the adjudication phase, a slight preference was expressed for annotations made by the first of the two annotators. Of the 328 items in disagreement, 200 (61%) were resolved in favour of the first annotator and 107 (33%) in favour of the second annotator. For the remaining 21 instances (6%), the adjudicator adopted the union of the two annotation sets.

Following adjudication, we are left with a data set in which 2079 sense annotations have been applied to 2038 instances, for an average of 1.02 senses per instance. This finding is in line with that of Henrich and Hinrichs (2014), who observe that the need to annotate more than one sense occurs infrequently. The special P/U senses were applied to 203 instances.

	Polysemy				
POS	1	2	3	4	Total
adjectives	48	3	0	0	51
nouns	33	11	6	1	51
verbs	28	17	5	1	51
total	109	31	11	2	153

Table 2: Number of lemmas in GLASS by part of speech and polysemy in GermaNet

6.2 Characterizing Lexical Substitutions

As mentioned in §3, the constructors of the GermEval data set had made a conscious decision not to control for polysemy in order to avoid biasing their selection of lemmas to any one sense inventory. Perhaps as a result, GLASS does not exhibit as wide a range of sense coverage as other sense-annotated data sets. Table 2 shows the frequency of the lemmas in GLASS by part of speech and polysemy in GermaNet 9.0. About half the verbs, two thirds of the nouns, and nearly all the adjectives have only a single sense listed in GermaNet. However, the average number of senses per lemma, 1.40, is still higher than GermaNet's overall average of 1.31.

We next undertake an investigation to determine the sort of lexical-semantic relations that hold between a disambiguated target and its substitutes. A similar study had been conducted by Kremer et al. (2014), though the sense annotations in their data set were automatically induced. Ours is therefore the first such study using manually applied sense annotations; it is also the first study using German-language data.

6.2.1 Substitute Coverage

We first consider GermaNet's coverage of the data set's 4224 unique *substitute types*—that is, the union of all words and phrases suggested by Cholakov et al.'s annotators, with duplicates removed. Of these types, only 3010 (71%) are found in GermaNet. Among the 1214 substitute types missing from GermaNet are many phrases or multiword expressions (38%), nominalizations of verbs which do occur in GermaNet (about 3%), and other derivations and compounds. There does not appear to be a great difference in lexical coverage for substitute types applied to items with successful versus unsuccessful sense annotations: 1081 of the 3887 unique substitute types applied to successfully sense-annotated items were not found in GermaNet (28%), as compared to 163 of the 667 types applied to the P/U items (24%).

6.2.2 Relating Targets and Substitutes

We next consider the semantic relations that link the successfully annotated target senses to their lexical substitutes. Recall that in GermaNet, words are grouped into structures known as synsets, where all words in the synset are synonymous. Synsets are in turn represented as vertices in a graph structure, with named semantic relations as the connecting edges. Table 3 shows the percentage of *substitute tokens* (*i.e.,* the individual words or phrases proposed as substitutes for each target, disregarding their frequency among annotators) which are synonyms, direct hypernyms, transitive hypernyms, direct hyponyms, or transitive hyponyms of any of its target's annotated senses. (The figures for transitive hypernyms and hyponyms exclude the direct hypernyms and hyponyms—that is, the target synset and the synset containing the substitute are endpoints on a path of length 2 or greater.) The table also shows the percentage of substitutes directly reachable by following any other type of semantic relation, the percentage of substitutes which exist in GermaNet but are not reachable from the target sense(s) via a path of uniform semantic relations, and the percentage of substitutes not covered by GermaNet at all.[5]

Relation	Adj.	Nouns	Verbs	Total
Synonym	7.5	6.6	4.6	5.9
Direct hypernym	7.1	7.5	6.5	6.9
Transitive hypernym	0.2	3.3	1.5	1.6
Direct hyponym	3.0	4.9	3.1	3.5
Transitive hyponym	1.5	0.7	0.8	0.6
Other direct relation	0.0	0.0	0.0	0.0
Otherwise reachable	60.4	58.9	71.2	65.4
Not in GermaNet	21.5	18.6	12.3	16.3

Table 3: Percentage of substitutes in successfully sense-annotated items in GLASS by their connection to the sense(s) through various semantic relations in GermaNet

From these statistics we can make a number of observations and comparisons to the English-language study of Kremer et al. (2014). First, the proportion of substitute tokens not in GermaNet is slightly lower than the proportion of substitute types not in GermaNet (16% *vs.* 24%). That is, of all substitute types in the data set, the annotators were more likely to apply those in GermaNet. Nonetheless, GermaNet's coverage of the substitutes in GLASS (84%) is significantly lower than WordNet's coverage of the substitutes in CoInCo (98%). Some of this difference must be due to how strictly each study's annotation guidelines discouraged the use of phrases and multiword expressions, which are largely absent from both WordNet and GermaNet. Around 6% of the GLASS substitutes are not in GermaNet because they are phrases or multiword expressions; the same figure for CoInCo cannot be more than 2%. The rest of the difference in substitute coverage may simply be a consequence of the size of the respective LSRs; WordNet 3.1 has about one and a third times the number of lemmas as GermaNet 9.0.

A second observation we can make is that the proportions of substitutes found in the synsets of the annotated senses and of those found in the synsets of the direct hypernyms are generally similar, while the proportion found in the synsets of transitive hypernyms is much lower. This is expected in light of the annotation instructions reported in Cholakov et al. (2014), which encouraged annotators to choose "a

[5] The numbers in each column of Table 3 may sum to slightly more than 100%, since a few words appear multiple times in the same hypernymy–hyponymy taxonomy. For example, in GermaNet the word *Öl* is its own hypernym, because it is a synonym of synset 40402 (petroleum oil) and also of the hypernym synset 48480 (a viscous liquid not miscible with water). This also holds for the English word *oil* in WordNet.

slightly more general word" only if there was no one word or phrase which perfectly fit the target. What is particularly surprising, however, is the sizeable proportion of substitutes found in the synsets of direct and transitive hyponyms. The reason this is surprising is that the annotation instructions did not make any provision for using more specific terms as substitutes. This anomaly was also observed in CoInCo, where direct and transitive hyponyms account for 7.5 and 3.0% of the substitutes, respectively.

Our third observation is that in no case is a substitute found in a synset directly related to the target by any semantic relation other than hypernymy or hyponymy. (GermaNet provides twelve such relation types, which are all (sub)classes of meronymy/holonymy, entailment, causation, and association.) This finding is also surprising, since it is not uncommon for meronyms or holonyms to serve as substitutes in German (Schemann, 2011, pp. 39*–43* [*sic*]). For example, as in English, the word *Person* ("person") can be substituted with its meronym *Kopf* ("head") in many contexts:

(1) Wir haben 8 € pro Person verdient.
[We earned €8 per person.]

(2) Wir haben 8 € pro Kopf verdient.
[We earned €8 per head.]

It is unclear whether or not semantic relations besides hypernymy and hyponymy produced any valid substitutes in CoInCo; Kremer et al. (2014) do not include them in their analysis.

Finally, we note that the majority of substitutes cannot be reached by following semantic relations of a single type. That is, some 60% of all substitutes exist as synonyms somewhere in the GermaNet graph, but are reachable from the target synset only by following semantic relations of at least two different types. This observation was also made for CoInCo, where 69% of the substitutes exist outside the target's hypernym/hyponym taxonomy.

6.2.3 Comparing Parasets to Synsets

Kremer et al. (2014) introduce the term *paraset* to refer to the set of substitutions produced for each target in its context, and investigate to what extent their parasets follow the boundaries of WordNet synsets. As their data set does not include manual sense annotations, they sense-annotate their targets heuristically by selecting the synset that has the greatest number of synonyms in common with the paraset. To overcome the lexical gap problem, they extend each synset's synonyms with those of its immediate hypernyms and hyponyms.

To measure the extent to which the parasets contain substitutes from a single synset, one can compute the *cluster purity* (Manning et al., 2008, §16.3). This metric, borrowed from information retrieval, measures the accuracy of each cluster with respect to its best matching gold class:

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j \left|\omega_k \cap c_j\right|,$$

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters, $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ is the set of classes, and N is the number of objects being clustered. Purity values range from 0 to 1, where 0 is the absence of purity and 1 is total purity. For our purposes, Ω is the set of parasets in GLASS, $\mathbb{C}$ is the set of synsets in GermaNet, and N is the total number of substitute tokens. Like Kremer et al. (2014), we consider only those substitutes that are found in the target's synsets or those of its hypernyms and hyponyms, as it would otherwise be unclear whether low purity implies substitutes from a mixture of senses (which is what we are trying to measure) or simply a large number of substitutes reachable via relations other than hypernymy and hyponomy (which we already confirmed above). By necessity we also disregard those instances which our annotators tagged as P/U or with more than one sense.

Our overall purity is 0.801; the first line of Table 4 shows purity values broken down by part of speech. These results are comparable to those of Kremer et al. (2014), who report purities of 0.812 for nouns and 0.751 for verbs. This is good evidence that our substitutes—or at least, the ones which are synonyms or direct hyper-/hyponyms of the target—tend to follow the boundaries of single GermaNet synsets.

6.2.4 Similarity Between Same-sense Parasets

In the previous section, we analyzed those substitutes found in the immediate semantic neighbourhood of the target sense. However, because the majority of our substitutes are found outside this neighbourhood, we now perform an investigation which includes these more distant relatives. In particular, we are interested in determining the similarity of parasets representing the same word sense.

Paraset similarity can be quantified as the number and proportion of their substitutes in the *common core*—that is, the intersection of all parasets for targets tagged with the same sense. As Table 4 shows, the parasets in our data set (again, excluding those for P/U and multiply tagged instances) have about 6.4 substitutes on average, with adjectives being slightly more substitutable and nouns slightly less. Most of these parasets—about 59%—have a non-empty common core. The average common core size across all parts of speech is slightly less than one. This means that about one sixth to one fifth of the parasets' substitutes are shared among all occurrences of the same target–sense combination.

Though it is reassuring that a common core exists more often than not, the fact that our same-sense parasets have more non-shared than shared substitutes is interesting. Part of the explanation for this is that some of the substitutes proposed by the annotators are highly context-specific, and do not apply to other instances even when used in the same word sense. For example, one of the contexts for *Athlet* ("athlete") is as follows:

(3) Seine eher mäßige schauspielerische Begabung rechtfertigte Weissmüller mit den Worten: „Das Publikum verzeiht meine Schauspielerei, weil es weiß, dass ich ein *Athlet* bin.“
[Weissmuller justified his rather modest acting talent by saying, "The public will forgive my acting because they know that I'm an *athlete*."]

Here the paraset was {*Wettkämpfer, Sportler, Muskelprotz, Olympionike, Herkules*}, but the common core included

Measure	Adj.	Nouns	Verbs	Total
cluster purity	0.774	0.795	0.824	0.801
mean paraset size	7.008	5.445	6.532	6.377
mean common core size	0.928	0.954	0.862	0.903
% common cores non-empty	62.963	72.727	42.254	58.639
% substitutes in common core	14.268	22.540	14.662	16.667

Table 4: Paraset purity and common core statistics for GLASS, by part of speech

only *Wettkämpfer* and *Sportler*. While the other three terms are plausible synonyms for this broad sense of *Athlet*, they would not necessarily fit every context. In particular, *Olympionike* ("Olympian") suggests that one of the annotators has exploited his or her real-world knowledge of the context's subject (in this case, Hollywood actor and competitive swimmer Johnny Weissmuller).

Another factor contributing to the low proportion of common-core substitutes is the sample size. As Kremer et al. (2014) observe, even six annotators cannot be expected to exhaust all possible substitutes for a given context. In fact, our common core statistics are only slightly lower than ones reported for CoInCo. In that data set, only about a quarter to a third of paraset substitutes were found in their respective common cores.

7 Conclusion and Future Work

In this paper, we have presented GLASS, a manually sense- and substitution-annotated German-language data set, and Ubyline, the annotation tool used to produce its sense annotations. Ubyline improves on the state of the art in sense annotation tools by supporting multiple lexical-semantic resources, including GermaNet, and by offering an ergonomic and intuitive mouse-driven user interface. The interface is optimized for the production of lexical sample data sets, and allows annotators to view the local semantic hierarchy without swapping between different displays.

One of Ubyline's more innovative features is its ability to record timestamps for all annotator activity. One possible direction for future work, then, would be to analyze the timing data we have recorded in the production of GLASS. This would reveal whether there are any correlations between annotation time and the various properties of the target word or its contexts. Not only could this help predict annotation time for future data sets, but it may also be useful for assessing text difficulty in a readability setting.

Our manually sense-annotated data set, GLASS, is unique in providing both sense and lexical substitution annotations for the same targets. Our intention in doing this was to enable the data set to be used for both *in vitro* and *in vivo* evaluations of word sense disambiguation systems. Though many of the lemmas in GLASS are monosemous in GermaNet, our data is still useful for intrinsic evaluations where systems must distinguish not only between senses provided by the inventory but the special "unassignable"/"proper name" tags that indicate a sense is missing from the inventory. GLASS has the further advantage of having the greatest lemma coverage of any fully manually sense-annotated data set for German. And unlike some other data sets which lack verbs or adjectives, it features an equal distribution across parts of speech (as well as lemma frequency). It is also the only WSD data set for German to have been published in full under a free content licence.

The two annotation layers in GLASS have enabled us to conduct the first known empirical study of the relationship between manually applied word senses and lexical substitutions. Contrary to expectations, we found that synonymy, hypernymy, and hyponomy are the only semantic relations directly linking targets to their substitutes. Moreover, the substitutes in the target's hypernymy/hyponomy taxonomy tend to closely align with the synonyms of a single synset in GermaNet. Despite this, these substitutes account for a minority of those provided by the annotators. Nearly two thirds of the substitutes exist somewhere in GermaNet but cannot be reached by traversing the target's hypernymy/hyponomy taxonomy, and a sixth of the substitutes are not covered by GermaNet at all. These findings could be used to inform the design of future automatic lexical substitution systems.

The results of our analysis accord with those of a previous study on English-language data, but where the sense annotations were induced from the substitution sets by a fully automatic process. From this we can draw a couple of tentative conclusions. First, the relations the two studies discovered between word senses and lexical substitutions may prove to be of a universal nature, holding for other data sets and languages. Second, we have gathered good (albeit indirect) evidence that lexical substitution data can be used as a knowledge source for automatic WSD. This finding suggests that training data annotated with respect to a fine-grained sense inventory such as WordNet could be produced semi-automatically, by deriving it from relatively cheap, manually applied lexical substitution annotations.

8 Acknowledgments

This work was supported by the DFG-funded project "Integrating Collaborative and Linguistic Resources for Word Sense Disambiguation and Semantic Role Labeling" (InCoRe, GU 798/9-1), by the DFG-funded research training group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES, GRK 1994/1), and by the Volkswagen Foundation as part of the Lichtenberg Professorship Program under grant № I/82806.

9 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.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of the 51st*

Annual Meeting of the Association for Computational Linguistics (ACL 2013), pages 1352–1362, August.

Bond, F., da Costa, L. M., and Lê, Tuấn Anh. (2015). IMI – A multilingual semantic annotation environment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (System Demonstrations) (ACL–IJCNLP 2015)*, pages 7–12, July.

Broscheit, S., Frank, A., Jehle, D., Ponzetto, S. P., Rehl, D., Summa, A., Suttner, K., and Vola, S. (2010). Rapid bootstrapping of word sense disambiguation resources for German. In *Proceedings of the 10th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2010)*, pages 19–27, September.

Cholakov, K., Biemann, C., Eckle-Kohler, J., and Gurevych, I. (2014). Lexical substitution dataset for German. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2524–2531.

Eckart de Castilho, R. and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, August.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, July.

C. Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY – A large-scale unified lexical-semantic resource. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590, April.

Hamp, B. and Feldweg, H. (1997). GermaNet – A lexical-semantic net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Henrich, V. and Hinrichs, E. (2010). GernEdiT – The GermaNet editing tool. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, May.

Henrich, V. and Hinrichs, E. (2013). Extending the TüBa-D/Z treebank with GermaNet sense annotation. In *Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013*, volume 8105 of *Lecture Notes in Artificial Intelligence*, pages 89–96. Springer.

Henrich, V. and Hinrichs, E. (2014). Consistency of manual sense annotation and integration into the TüBa-D/Z treebank. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 62–74, December.

Henrich, V., Hinrichs, E., and Suttner, K. (2012a). Automatically linking GermaNet to Wikipedia for harvesting corpus examples for GermaNet senses. *Journal for Language Technology and Computational Linguistics*, 27(1):1–19.

Henrich, V., Hinrichs, E., and Vodolazova, T. (2012b). 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 2012)*, pages 387–396, April.

Henrich, V., Hinrichs, E., and Vodolazova, T. (2014). Aligning GermaNet senses with Wiktionary sense definitions. In *Proceedings of the 5th Language and Technology Conference (LTC 2011)*, number 8387 in Lecture Notes in Artificial Intelligence, pages 329–342. Springer.

Henrich, V. (2015). *Word Sense Disambiguation with GermaNet: Semi-Automatic Enhancement and Empirical Results*. Ph.D. thesis, Eberhard Karls Universität Tübingen, January.

Kremer, G., Erk, K., Padó, S., and Thater, S. (2014). What substitutes tell us – Analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 540–549, April.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

McCarthy, D. (2002). Lexical substitution as a task for WSD evaluation. In *Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 89–115, July.

Mihalcea, R. and Chklovski, T. (2003). Open Mind Word Expert: Creating large annotated data collections with Web users' help. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, pages 53–60, April.

Miller, T. and Turković, M. (2016). Towards the automatic detection and identification of English puns. *European Journal of Humour Research*, 4(1), January.

Miller, T., Benikova, D., and Abualhaija, S. (2015). GermEval 2015: LexSub – A shared task for German-language lexical substitution. In *Proceedings of* GermEval 2015: LexSub, pages 1–9, September.

Passonneau, R. J., Salleb-Aouissi, A., and Ide, N. (2009). Making sense of word sense variation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW 2009)*, pages 2–9, June.

Raileanu, D., Buitelaar, P., Vintar, S., and Bay, J. (2002). Evaluation corpora for sense disambiguation in the medical domain. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 609–612.

Schemann, H. (2011). *Deutsche Idiomatik: Wörterbuch der deutschen Redewendungen im Kontext*. Walter de Gruyter, Berlin/Boston, MA, 2nd edition.

Tan, L. and Bond, F. (2011). Building and annotating the linguistically diverse NTU-MC (NTU-Multilingual Corpus). In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 362–371, December.