# GermEval 2015

# LexSub

Proceedings of the First Workshop on German Lexical Substitution

Compilation and design © 2015 UKP Lab, Technische Universität Darmstadt.
All articles © 2015 their respective authors.

This volume and its individual articles are licensed under the
Creative Commons Attribution 4.0 International (CC BY 4.0) licence:
https://creativecommons.org/licenses/by/4.0/

# GermEval 2015: LexSub
# Organizing Committee

**Sallam Abualhaija**
Institute of Computer Technology
Technische Universität Hamburg-Harburg

**Darina Benikova**
Research Training Group AIPHES
Department of Computer Science
Technische Universität Darmstadt

**Chris Biemann**
Language Technology Group
Department of Computer Science
Technische Universität Darmstadt

**Judith Eckle-Kohler**
Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt

**Iryna Gurevych**
Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt

**Tristan Miller**
Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt

# Workshop Program

**Tuesday, 29 September 2015**

| | |
|---|---|
| **13:45–13:55** | **Opening address** |
| **14:00–15:30** | **Invited talk**<br>*Lexical Substitution: From a Testbed for Natural Language Understanding to a Practical Technology*<br>György Szarvas |
| **15:30–16:00** | **Coffee break** |
| **16:00–17:30** | **Main session** |
| | *GermEval 2015: LexSub – A Shared Task for German-language Lexical Substitution*<br>Tristan Miller, Darina Benikova, and Sallam Abualhaija |
| | *Delexicalized Supervised German Lexical Substitution*<br>Gerold Hintz and Chris Biemann |
| | *Lexical Substitution Using Deep Syntactic and Semantic Analysis*<br>Luchezar Jackov |
| **17:30–18:00** | **Discussion session** |

# Invited Talk:
# Lexical Substitution: From a Testbed for Natural Language Understanding to a Practical Technology

**György Szarvas**

Amazon Development Center Germany GmbH

## Abstract

Lexical substitution has been an area of intensive research for much of the past decade. A significant portion of research interest in the task has centered around using lexical paraphrasing as a testbed to evaluate (vector-based) semantic models. More recently there is growing interest in lexical substitution as a standalone task. This is fueled by both the progress of the state of the art in solving this task and by the increasing number of practical applications for which lexical substitution is a core technology.

In this talk I will provide a brief overview of the recent advances in lexical substitution, and present our work that aimed to improve the accuracy of lexical substitution by leveraging the power of supervised models (while preserving the ability to address the problem in an open vocabulary setting). In the second part of the talk, I will present some practical applications that can benefit from an accurate lexical substitution system and discuss some aspects of the task (such as coverage, evaluation metrics, etc.) that seem to be important from an application perspective.

## Biography

György Szarvas is a machine learning scientist at Amazon in Berlin, Germany. His research interests include lexical semantics, information extraction and the application of machine learning techniques to NLP. He received his Ph.D. degree in computer science in 2008 from the University of Szeged, Hungary where he worked on domain and language independent named entity recognition, and uncertainty detection in biomedical texts.

From 2009–2012 he was a senior researcher at UKP Lab, Technische Universität Darmstadt, working on lexical semantics (lexical substitution and detection of uncertain statements), and learning to rank for information retrieval.

In 2012 he joined Nuance Communications in Aachen as a research engineer working on information extraction from medical texts for automated question answering. Since 2013 he has worked at Amazon Berlin as member of the NLP team and works on improving the quality and extracting valuable information from user generated content (customer reviews).

# Table of Contents

# GermEval 2015: LexSub – A Shared Task for German-language Lexical Substitution

**Tristan Miller**
Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt

**Darina Benikova**
Research Training Group AIPHES
Department of Computer Science
Technische Universität Darmstadt

**Sallam Abualhaija**
Institute of Computer Technology
Technische Universität Hamburg-Harburg

## Abstract

Lexical substitution is a task in which participants are given a word in a short context and asked to provide a list of synonyms appropriate for that context. This paper describes GermEval 2015: LexSub, the first shared task for automated lexical substitution on German-language text. We describe the motivation for this task, the evaluation methods, and the manually annotated data set used to train and test the participating systems. Finally, we present an overview and discussion of the participating systems' methodologies, resources, and results.

## 1 Introduction

Word sense disambiguation, or WSD (Agirre and Edmonds, 2007)—the task of determining which of a word's senses is the one intended in a particular context—has been a core research problem in computational linguistics since the very inception of the field. Approaches to WSD system evaluation can be categorized as *intrinsic* (or *in vitro*) or *extrinsic* (*in vivo*) (Ide and Véronis, 1998). In the former, the assessment is performed independently of any particular natural language processing application. Rather, evaluators directly compare the automatically produced sense assignments with a manually annotated gold standard (Palmer et al., 2007). In extrinsic evaluation, however, systems are scored according to their contribution to a dedicated NLP task, such as machine translation (Carpuat and Wu, 2005a,b; Chan et al., 2007; Carpuat and Wu, 2007) or information retrieval (Clough and Stevenson, 2004; Schütze and Pedersen, 1995; Sanderson, 1994; Zhong and Ng, 2012).

Most published WSD evaluations to date, such as those in the Senseval and SemEval workshop series, have been of the intrinsic variety. However, it is widely agreed that extrinsic evaluations are preferable, since the usual point of computational WSD is to support real-world NLP applications. The idea of using lexical substitution for *in vivo* WSD evaluation was proposed as far back as 2002 (McCarthy, 2002) and has led to a number of English, Italian, and crosslingual evaluation competitions since then (McCarthy and Navigli, 2007; Toral, 2009; Mihalcea et al., 2010). Until now, however, no one has conducted a rigorous evaluation of lexical substitution systems on German-language text. In this paper, we describe and present the results of GermEval 2015: LexSub, the scientific community's first shared task for German-language lexical substitution.

The remainder of this paper is structured as follows: §2 reviews the task of lexical substitution and the methodologies used to evaluate the performance of lexical substitution systems, §3 describes the data set used to train and test the systems participating in our task, and §4 describes the lexical-semantic resources made available to the participants and employed by some of the systems and baselines. In §§5 and 6 we briefly describe these systems and baselines, respectively, and in §7 we present and discuss their results on the test data set. Finally, we wrap things up in §8 with some general observations.

## 2 Task definition

Lexical substitution is the task of identifying appropriate substitutes for a target word in a given context. For example, consider the following two German-language contexts (abridged from Cholakov et al. (2014)) containing the word *Erleichterung*:

(1) *In der Legislaturperiode 1998–2002 wurden einige Reformen des Staatsbürgerschaftsrechts bezüglich der* ***Erleichterung*** *von Einwanderung verabschiedet.*

(In the legislative period of 1998–2002 a few reforms on citizenship law concering the **easing** of immigration were passed.)

(2) *Vor allem auf dem Lande war die Umstellung aber schwer durchsetzbar und die* ***Erleichterung*** *groß, als 1802 der Sonntagsrhythmus und 1805 der vorrevolutionäre Kalender insgesamt wieder eingeführt wurden.*

(The change was particularly difficult to enforce in the countryside, and there was great **relief** when in 1802 the Sunday routine and in 1805 the pre-revolutionary calendar were reintroduced.)

The word *Förderung* (meaning "facilitation") would be an appropriate substitute for *Erleichterung* (meaning "easing") in the first context, whereas the word *Freude* (meaning "delight") would not be. Conversely, *Freude* would indeed be a valid substitute for *Erleichterung* (meaning "relief") in the second context, whereas *Förderung* would not be.

Lexical substitution is a relatively easy task for humans, but potentially very challenging for machines because it relies—explicitly or implicitly—on word sense disambiguation, a longstanding core problem in computational linguistics. In fact, lexical substitution was originally conceived as a method for evaluating word sense disambiguation systems which is independent of any one sense inventory. However, it also has a number of uses in real-world NLP tasks, such as text summarization, question answering, paraphrase acquisition, text categorization, information extraction, text simplification, lexical acquisition, and text watermarking.

Evaluation of automated lexical substitution systems is effected by applying them on a large number of word–context combinations (*items* or *instances*) and then comparing the substitutions they propose to those made by human annotators. There are various scoring methodologies which have been used in past lexical substitution tasks. The following list briefly describes the ones employed in our task; for details of their derivation and precise computation the reader is referred to the cited papers.

**Best** (McCarthy and Navigli, 2009) allows a system to propose as many substitutes as it wishes for each item, but considers the first proposed substitute to be its "best guess". This methodology uses the following metrics:

**Recall (R)** scores each item by finding the average human annotator response frequency of the system's substitutes and dividing by the number of system substitutes. The scores for all items are then summed and divided by the total number of items in the data set.

**Precision (P)** is the same as recall, except that items for which the system declined to propose any substitutes are disregarded.

**Mode recall (Mode R)** is the number of times the system's "best guess" corresponded to the one substitute most commonly chosen by the human annotators, divided by the number of items with such a human-annotated substitute.

**Mode precision (Mode P)** is the same as mode recall, except that items for which the system declined to propose any substitutes are disregarded.

**Out-of-ten (OOT)** (McCarthy and Navigli, 2009) allows a system to propose up to ten substitutes for each item, though the order of these is not significant. The following scoring metrics are used:

**Recall (R)** is the same as the *best* recall metric, except that the credit for each correct substitute is not divided by the number of proposed substitutes.

**Precision (P)** is the same as the *best* precision metric, except that the credit for each correct substitute is not divided by the number of proposed substitutes.

**Mode recall (Mode R)** is the number of times the one substitute most commonly chosen by the human annotators occurred among the system's substitutes, divided by the number of items for which there was a single most frequent human-annotated substitute.

**Mode precision (Mode P)** is the same as mode recall, except that items for which the system declined to propose any substitutes are disregarded.

**Generalized average precision (GAP)** (Kishida, 2005) allows a system to propose a ranked list of substitutes and then assesses the quality of the entire ranked list. It is believed to be superior to *OOT* because of its sensitivity to the relative position of correct and incorrect candidates in the ranking.

## 3 Data set

For our training and test data, we use the German-language lexical substitution data set produced by Cholakov et al. (2014). The full data set consists of 2040 context sentences from the German edition of Wikipedia, each containing one target word. There are 153 unique target words, equally distributed across parts of speech (nouns, verbs, and adjectives) and three frequency groups according to the lemma frequency list of the German WaCky corpus (Baroni et al., 2009). There are ten context sentences for each noun and adjective target, and twenty for each verb. Two hundred of the sentences were annotated by four professional human annotators, and the remainder by one professional annotator and five additional annotators recruited via crowdsourcing. About half of this data (26 nouns, 26 verbs, and 26 adjectives in 1040 sentence contexts) forms the training set, which was made available to participants in full in advance of the task. The remainder forms the test set, which (excluding the list of substitutions) was given to the participants at the beginning of the task.

This German data set is similar in size and scope to past English and Italian data sets. The SemEval-2007 lexical substitution data set consists of 2010 sentences (ten sentences for each of 201 unique target words) and the EVALITA 2009 data contains 2310 sentences (also with ten sentences per word). In contrast to the English and Italian data sets, the Cholakov et al. (2014) data has a greater emphasis on verbs, and contains no adverbs since the distinction between adverbs and adjectives is less pronounced in German.

We have now published the entire data set, including the human-provided substitutions, under the Creative Commons Attribution-ShareAlike license.[1] This is, to our knowledge, the only published data set which makes possible the evaluation of WSD systems with an arbitrary sense inventory. (Existing collections of sense-annotated German text, such as WebCAGe (Henrich et al., 2012) and TüBa-D/Z (Henrich and Hinrichs, 2013), are all tied to GermaNet.)

The format of the files in the data set corresponds to that of lexical substitution tasks in other languages (McCarthy and Navigli, 2007; Toral, 2009). There are two types of files:

1. XML files containing single-sentence instances enclosed in `instance` and `context` elements. Within each instance, the target word is enclosed in a `head` element. Instances with the same target lemma are grouped together in a `lexelt` element. The `lexelt` elements are grouped together in a top-level `corpus` element. The entire format is illustrated in Figure 1.

2. Delimited *gold* files which are cross-referenced to the XML files and which contain the gold-standard substitutions. Each line has the format

   ```
   lexelt id :: subs
   ```

   where

   `lexelt` is the unique identifier for the target lemma, corresponding to the `item` attribute of the `lexelt` element in the XML file;

   `id` is the unique identifier for the instance, which matches the `id` attribute of the `instance` element; and

   `subs` is a semicolon-delimited list of lemmatized substitutes. Each substitute is followed by a space and its corresponding frequency count (indicating the number of annotators who provided that substitute).

   The *gold* file line corresponding to the instance shown in Figure 1 is shown in Figure 2.

## 4 Resources

We made available to all participants a number of language resources supporting the task of lexical substitution:

**GermaNet** (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) is a lexical-semantic network that relates German-language nouns, verbs, and adjectives. It is the analogue of WordNet (Fellbaum, 1998) and ItalWordNet (Roventini et al., 2000) used in past English

[1] `https://www.ukp.tu-darmstadt.de/data`

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE corpus SYSTEM 'lexsub.dtd'>
<corpus lang="de">
  <lexelt item="Monarch.n">
    <instance id="Monarch_1">
      <context>
        Dies war die letzte britische Regierung, die ein <head>Monarch</head>
        ohne Mehrheit im Unterhaus ernannte, und scheiterte schon im April 1835.
      </context>
    </instance>
    ⋮
  </lexelt>
  ⋮
</corpus>
```

Figure 1: Format of the data set's XML files

```
Monarch.n Monarch_1 :: König 3; Herrscher 2; Adliger 1; Staatsoberhaupt 1;
```

Figure 2: Sample line from a *gold* file

| Resource | Senses | Synsets |
|---|---|---|
| WordNet 2.1 | 207 016 | 117 597 |
| WordNet 3.0 | 206 941 | 117 659 |
| ItalWordNet | ca. 130 000 | ca. 80 000 |
| GermaNet 8.0 | 111 361 | 84 584 |
| GermaNet 9.0 | 121 810 | 93 246 |
| GermaNet 10.0 | 131 814 | 101 371 |

Table 1: Comparison of language resources used for lexical substitution

and Italian lexical substitution tasks, respectively. All three wordnets group word–sense pairs (*lexical units* or *senses*) expressing the same concept into structures called *synsets*.

The standalone version of GermaNet offered to GermEval 2015: LexSub participants was GermaNet 10.0, though they also had the choice of using GermaNet 9.0 as part of UBY (see below). The baselines described in §6 use GermaNet 8.0.

Table 1 shows the number of senses and synsets for the versions of WordNet, ItalWordNet, and GermaNet used in the current and past lexical substitution tasks.

**UBY** (Gurevych et al., 2012) is a large-scale lexical-semantic resource which links information from several expert- and collaboratively constructed resources for English and German. The linked resources include GermaNet 9.0, WordNet 3.0, and the English and German versions of Wikipedia and Wiktionary.

**JoBimText** (Biemann and Riedl, 2013) is an automatically induced resource for German by means of distributional semantics. Distributional thesauri, as well as distributional features of words, are provided as a RESTful API and as a database. These features were demonstrated to be beneficial for lexical substitution by Szarvas et al. (2013).

## 5 Participating systems

GermEval 2015: LexSub saw participation from two systems, from Hintz and Biemann (2015) and Jackov (2015), though as the former is connected with one of the task organizers, it was entered non-competitively.

Hintz and Biemann use a supervised delexicalized approach adapted from previous work on English-language lexical substitution by Szarvas et al. (2013). They made use of Wiktionary and GermaNet (via UBY) and the JoBimText distributional thesauri, as well as the online lexical resources Woxikon, Duden, and Leipzig Wortschatz. They employ a maximum entropy classifier, regarding the task as a binary classification problem on whether any given substitution fits or does not fit the context. In addition to the semantic resource features, they make use of frequency, co-occurrence, and embedding features.

Jackov applies a deep semantic and syntactic approach relying on machine translation techniques.

| System | Best P | Best R | Best Mode P | Best Mode R | OOT P | OOT R | OOT Mode P | OOT Mode R | GAP |
|---|---|---|---|---|---|---|---|---|---|
| RandomSense | 7.40 | 7.40 | 15.13 | 15.13 | 12.53 | 12.53 | 23.45 | 23.45 | 9.54 |
| TopRankedSynonym | 10.04 | 10.04 | 19.82 | 19.82 | 15.21 | 15.21 | 27.99 | 27.99 | 12.25 |
| WeightedSense | 7.50 | 7.50 | 13.46 | 13.46 | **20.54** | **20.54** | **35.55** | **35.55** | 14.28 |
| Hintz and Biemann[a] | **11.20** | **11.10** | **24.28** | **24.21** | 19.49 | 19.31 | 33.99 | 33.89 | **15.96** |
| Jackov | 6.73 | 6.45 | 13.36 | 12.86 | 20.14 | 19.32 | 33.18 | 31.92 | 11.26 |

[a] System co-authored by one of the task organizers

Table 2: Baseline and system results for the *best*, *OOT*, and *GAP* metrics

Apart from the English WordNet, the author employs a custom-built machine translation system and a dependency relation knowledge base. The approach first disambiguates the input text by tentatively mapping the lemmatized German words to concepts represented by WordNet synsets. Each parsing hypothesis is scored with reference to a knowledge base of dependency relations; the synonyms and hypernyms of the target concept in the highest-scoring parsing hypothesis are taken as the substitution candidates.

## 6 Baselines

In addition to the dedicated lexical substitution systems described in the previous section, we implemented three simple baselines, at least two of which have been used in previous lexical substitution tasks:

**RandomSense** selects a random sense of the target word from GermaNet and returns its synonyms, followed by its hypernyms, in the same order as retrieved from the GermaNet API.

**TopRankedSynonyms** (McCarthy and Navigli, 2009) builds a list of substitutes in the following order:

1. Synonyms from the first synset of the target word, ranked according to their frequency in a large corpus.
2. Synonyms from the hypernyms (verbs and nouns) or closely related classes (adjectives) from the first synset, ranked according to their frequency in a large corpus.
3. Synonyms from all other synsets of the target word, ranked according to their frequency in a large corpus.
4. Synonyms from the hypernyms (verbs and nouns) or closely related classes (adjectives) of all other synsets of the target word, ranked according to their frequency in a large corpus.

**WeightedSense** (Toral, 2009) uses multiple lexical-semantic resources to build the list of candidates. In our case, we use GermaNet and Wiktionary to extract all synonyms and hypernyms of the target word. Synonyms are given a weight of 3, and hypernyms a weight of 1. If a substitute is extracted more than once (i.e., from different synsets or resources), the weights are summed. The list is then ordered by descending weight.

## 7 Results

Table 2 shows the baseline and participating systems' results for the various *best*, *OOT*, and *GAP* metrics, represented as percentages, on the test set. For each metric, the score for the best-performing system or baseline is set in boldface. Unsurprisingly, RandomSense is the worst-performing baseline. TopRankedSynonym performs best among the baselines by the *best* methodology and the WeightedSense baseline performs best according to both the *OOT* and the *GAP* methodologies.

With respect to the participants' systems, we observe that Hintz and Biemann's entry greatly outperforms Jackov's on the *best* and *GAP* metrics. In fact, the latter fails to beat even the baseline systems in *best*, pointing to the lack of an appropriate substitute ranking scheme. However, for *OOT*,

| | | Best | | | | OOT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mode | Mode | | | Mode | Mode | |
| | System | P | R | P | R | P | R | P | R | GAP |
| adjectives | RandomSense | 7.31 | 7.31 | 17.18 | 17.18 | 16.01 | 16.01 | 35.58 | 35.58 | 11.87 |
| | TopRankedSynonym | 9.63 | 9.63 | 23.31 | 23.31 | 16.01 | 16.01 | 35.58 | 35.58 | 13.85 |
| | WeightedSense | 6.10 | 6.10 | 11.66 | 11.66 | 20.73 | **20.73** | **42.33** | **42.33** | 15.06 |
| | Hintz and Biemann[a] | **14.20** | **13.69** | **36.02** | **35.58** | **21.29** | 20.53 | 42.24 | 41.72 | **18.86** |
| | Jackov | 4.58 | 4.07 | 10.20 | 9.20 | 16.94 | 15.04 | 29.93 | 26.99 | 7.35 |
| nouns | RandomSense | 8.42 | 8.42 | 14.02 | 14.02 | 16.79 | 16.79 | 23.78 | 23.78 | 12.46 |
| | TopRankedSynonym | **12.73** | **12.73** | 20.73 | 20.73 | 18.56 | 18.56 | 26.22 | 26.22 | 15.96 |
| | WeightedSense | 8.80 | 8.80 | 9.76 | 9.76 | 24.43 | 24.43 | 35.37 | 35.37 | 16.32 |
| | Hintz and Biemann[a] | 11.11 | 11.11 | **21.95** | **21.95** | **26.38** | **26.38** | **39.63** | **39.63** | **19.61** |
| | Jackov | 10.58 | 10.49 | 17.90 | 17.68 | 21.61 | 21.44 | 31.48 | 31.10 | 14.62 |
| verbs | RandomSense | 6.93 | 6.93 | 14.67 | 14.67 | 8.65 | 8.65 | 17.37 | 17.37 | 6.92 |
| | TopRankedSynonym | 8.89 | 8.89 | 17.66 | 17.66 | 13.14 | 13.14 | 25.15 | 25.15 | 9.59 |
| | WeightedSense | 7.55 | 7.55 | 16.17 | 16.17 | 18.50 | 18.50 | 32.34 | 32.34 | **12.87** |
| | Hintz and Biemann[a] | **9.80** | **9.80** | **19.76** | **19.76** | 15.17 | 15.17 | 27.25 | 27.25 | 12.69 |
| | Jackov | 5.75 | 5.62 | 12.54 | 12.28 | **20.86** | **20.40** | **35.47** | **34.73** | 11.53 |

[a] System co-authored by one of the task organizers

Table 3: Baseline and system results for the *best*, *OOT*, and *GAP* metrics, by part of speech

| | Best | | | | OOT | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Mode | Mode | | | Mode | Mode |
| System | P | R | P | R | P | R | P | R |
| Yuret | **12.90** | **12.90** | 20.65 | 20.65 | 46.15 | 46.15 | 61.30 | 61.30 |
| Hassan et al. | 12.77 | 12.77 | **20.73** | **20.73** | 49.19 | 49.19 | **66.26** | **66.26** |
| Giuliano et al. | 6.95 | 6.94 | 20.33 | 20.33 | **69.03** | **68.90** | 58.54 | 58.54 |
| TopRankedSynonym | 9.95 | 9.95 | 15.28 | 15.28 | 29.70 | 29.35 | 40.57 | 40.57 |

Table 4: Top-performing baseline and system results for SemEval-2007

| | Best | | | | OOT | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Mode | Mode | | | Mode | Mode |
| System | P | R | P | R | P | R | P | R |
| Basile and Semeraro | 8.16 | 7.18 | 10.58 | 10.58 | **41.46** | **36.50** | **47.23** | **47.23** |
| WeightedSense[a] | **10.86** | **9.06** | **13.94** | **13.94** | 23.00 | 19.20 | 26.97 | 26.97 |
| WeightedSense[b] | 9.71 | 8.19 | 13.16 | 13.16 | 27.52 | 23.23 | 37.24 | 32.39 |

[a] CLIPS only
[b] CLIPS and ItalWordNet

Table 5: Top-performing baseline and system results for EVALITA 2009

Jackov's performance is on par with, and occasionally exceeds, that of Hintz and Biemann. Neither system was able to beat the WeightedSense baseline for any of the metrics in *OOT*.

When broken down by part of speech (see Table 3), we observe that scores of the best-performing systems are generally higher for adjectives and nouns, but lower for verbs. It has long been known that verbs are the hardest category of words to process in traditional WSD (Agirre and Stevenson, 2007); it seems this holds for lexical substitution as well. The part-of-speech breakdown also allows us to see that some systems perform better, relative to the others, for different word categories. Of particular note is the TopRankedSynonym baseline's high precision and recall scores for nouns in the *best* methodology, and Jackov's outstanding performance on verbs across all *OOT* metrics. An optimal lexical substitution system may therefore benefit from adapting its strategy according to the target's part of speech.

We also performed an analysis of the relationship between system scores and target word frequency using the Pearson product-moment correlation coefficient. For each combination of system and scoring metric we observed only a negligible negative correlation ($-0.188 \leq r \leq -0.003$). The correlation between system scores and target word polysemy was also computed; this was weak at best ($-0.219 \leq r \leq -0.039$).

### 7.1 Comparison to SemEval and EVALITA

As previously mentioned, the English SemEval-2007 and Italian EVALITA 2009 shared tasks use similar data sets to our own, as well as some of the same baselines and evaluation methodologies. It is therefore interesting to compare the results of these baselines, and those of their top-performing systems, to our own.

Table 4 shows the results of the best-performing SemEval-2007 system for each of the *best* and *OOT* metrics (Yuret, 2007; Hassan et al., 2007; Giuliano et al., 2007). Also shown there are results for their TopRankedSynonym baseline, which uses WordNet 2.1. Again, for each column the best-performing system or baseline is set in boldface. We observe that the GermaNet-based TopRankedSynonyms baseline performs slightly better than its English counterpart for all the *best* metrics, but significantly worse for all the *OOT* metrics. As in GermEval 2015: LexSub, at least one participating system was able to beat the TopRankedSynonym baseline for any given metric. However, the relative improvement over the baseline was dramatically higher in the English-language task (29.6% to 132.4% in SemEval as compared to 10.2% to 37.6% in GermEval).

Table 5 shows a corresponding results table for the EVALITA 2009 shared task. Here we report scores for two implementations of the WeightedSense baseline; the first uses only the CLIPS lexical-semantic resource (Ruimy et al., 2002), whereas the second, like our own WeightedSense, uses two resources: CLIPS and ItalWordNet. The top-performing participating system here was one submitted by Basile and Semeraro (2009). As in GermEval, in EVALITA scores for the WeightedSense baseline frequently exceeded those of the participating systems. Interestingly, the circumstances under which this occurred were quite different: in GermEval, WeightedSense bested the participating systems for most of the *OOT* metrics, whereas in EVALITA, it was the *best* metrics in which the baseline excelled. German systems may be performing worse due to a lack of lexical coverage in GermaNet, or possibly, as Hintz and Biemann (2015) speculate, because its graph structure makes its lexical items harder to discover.

## 8 Concluding remarks

In this paper we have introduced GermEval 2015: LexSub, the first lexical substitution task using German text, and presented the results of three baselines and two participating systems. Due to the very low number of participating systems compared with previous lexical substitution tasks in other languages, it is difficult to draw any firm conclusions concerning the efficacy of the different approaches. On the one hand, one of the systems has shown that techniques proven to work well for English-language lexical substitution can work well for German too. But on the other hand, the second system, taking a completely novel approach, had comparable performance much of the time, and the rest of the time seemed to be held back only by its substitute ranking criteria.

Compared with previous lexical substitution tasks, our absolute scores in the *best* metrics were in about the same range, though relative to the baselines they were much lower than in SemEval and much higher than in EVALITA. Unlike in the English and Italian tasks, our participants' systems had trouble beating the baselines for *OOT*, suggesting that the problem may be lack of lexical coverage in German language resources, or the systems' inability to exploit this coverage.

## Acknowledgments

The GermEval 2015: LexSub shared task 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), the BMBF-funded Common Language Resources and Technologies Infrastructure (CLARIN, F-AG7), and DFG-funded research training group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES, GRK 1994/1).

## References

Eneko Agirre and Philip Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech, and Language Technology*. Springer, 2007. ISBN 978-1-4020-6870-6.

Eneko Agirre and Mark Stevenson. Knowledge sources for WSD. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech, and Language Technology*. Springer, 2007. ISBN 978-1-4020-6870-6.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009. ISSN 1574-020X.

Pierpaolo Basile and Giovanni Semeraro, Baroni. UNIBA @ EVALITA 2009 lexical substitution task. In *Proceedings of EVALITA 2009*, 2009.

Chris Biemann and Martin Riedl. Text: Now in 2D! A framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95, 2013. ISSN 2299-856X.

Marine Carpuat and Dekai Wu. Evaluating the word sense disambiguation performance of statistical machine translation. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP 2005)*, 2005a.

Marine Carpuat and Dekai Wu. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43th Annual Meeting of the Association of Computational Linguistics (System Demonstrations) (ACL 2005)*, 2005b.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP–CoNLL 2007)*, June 2007.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, 2007.

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

Paul Clough and Mark Stevenson. Evaluating the contribution of EuroWordNet and word sense disambiguation to cross-language retrieval. In *Proceedings of the 2nd International Conference of the Global WordNet Association (GWC 2004)*, pages 97–105, 2004.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998. ISBN 978-0-262-06197-1.

Claudio Giuliano, Alfio Gliozzo, and Carlo Strapparava. FBK-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 145–148, 2007.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 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 2012.

Birgit Hamp and Helmut Feldweg. 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, 1997.

Samer Hassan, Andras Csomai, Carmen Banea, Ravi Sinha, and Rada Mihalcea. UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 410–413, 2007.

Verena Henrich and Erhard Hinrichs. GernEdiT – The GermaNet editing tool. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, May 2010.

Verena Henrich and Erhard Hinrichs. Extending the TüBa-D/Z treebank with GermaNet sense annotation. In Iryna Gurevych, Chris Biemann, and Torsten

Zesch, editors, *Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013*, volume 8105 of *Lecture Notes in Artificial Intelligence*, pages 89–96. Springer, 2013.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 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 2012.

Gerold Hintz and Chris Biemann. Delexicalized supervised German lexical substitution. In *Proceedings of GermEval 2015: LexSub*, pages 11–16, September 2015.

Nancy Ide and Jean Véronis. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998. ISSN 0891-2017.

Luchezar Jackov. Lexical substitution using deep syntactic and semantic analysis. In *Proceedings of GermEval 2015: LexSub*, pages 17–20, September 2015.

Kazuaki Kishida. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Technical Report NII-2005-014E, National Institute of Informatics, Tokyo, Japan, October 2005.

Diana McCarthy. Lexical substitution as a task for WSD evaluation. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 109–115, 2002.

Diana McCarthy and Roberto Navigli. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, 2007.

Diana McCarthy and Roberto Navigli. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159, 2009. ISSN 1574-020X.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. SemEval-2010 Task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, pages 9–14, July 2010.

Martha Palmer, Hwee Tou Ng, and Hoa Trang Dang. Evaluation of WSD systems. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech, and Language Technology*. Springer, 2007. ISBN 978-1-4020-6870-6.

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. ItalWordNet: A large semantic database for Italian. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 783–790, 2000.

Nilda Ruimy, Monica Monachini, Raffaella Distante, Elisabetta Guazzini, Stefano Molino, Marisa Ulivieri, Nicoletta Calzolari, and Antonio Zampolli. Clips, a multi-level Italian computational lexicon: A glimpse to data. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 2002.

Mark Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1994)*, pages 142–151, 1994.

Hinrich Schütze and Jan O. Pedersen. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1995.

György Szarvas, Chris Biemann, and Iryna Gurevych. Supervised all-words lexical substitution using delexicalized features. In *Proceedings of the 10th Conference of the North American Chapter of the Association for Computational Linguistics and the 18th Human Language Technologies Conference (NAACL–HLT 2013)*, pages 1131–1141, 2013.

Antonio Toral. The lexical substitution task at EVALITA 2009. In *Proceedings of EVALITA 2009*, 2009.

Deniz Yuret. Ku: Word sense disambiguation by substitution. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 207–214, 2007.

Zhi Zhong and Hwee Tou Ng. Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics (ACL 2012)*, pages 273–282, July 2012.

# Delexicalized Supervised German Lexical Substitution

**Gerold Hintz** and **Chris Biemann**
Research Training Group AIPHES / FG Language Technology
Department of Computer Science, Technische Universität Darmstadt
{hintz,biem}@lt.informatik.tu-darmstadt.de

## Abstract

We address the German lexical substitution task, which requires retrieving a ranked list of meaning-preserving substitutes for a given target word within an utterance. With *GermEval-2015: LexSub*[1], this challenge is posed for the first time using German language data. In this work we build upon the existing state of the art for English lexical substitution, employing a delexicalized supervised system. In adapting the existing approach, we consider in particular the available lexical resources for German and evaluate their suitability to the task at hand. We report first results on German lexical substitution and observe a similar performance as English systems evaluated on the SemEval dataset.

## 1 Introduction

Lexical substitution is a special form of contextual paraphrasing which aims to predict substitutes for a target word instance within a sentence. This implicitly addresses the problem of resolving the ambiguity of polysemous terms. In contrast to Word Sense Disambiguation (WSD) this is achieved without requiring a predefined inventory of senses. A vector of substitute words for a given target can be regarded as an alternative contextualized meaning representation that can be used in similar downstream tasks such as Information Retrieval or Question Answering. In contrast to WSD, lexical substitution systems are not limited by the coverage or granularity of the underlying sense inventory, and is still applicable to languages in which no such resource is available at all. As a result, lexical substitution systems have become very popular for evaluating context-sensitive lexical inference since the introduction of the first *SemEval-2007 lexical substitution* task (McCarthy and Navigli, 2007). Whereas this and earlier variants of this task were posed without any training data and a relatively small evaluation set of a few thousand instances, later datasets were scaled up by the use of crowdsourcing, containing nearly 24k sentences with substitutes for a lexical sample of 1012 frequent nouns (Biemann, 2013). With *GermEval 2015*, German lexical substitution data (Cholakov et al., 2014) is provided for the first time. The dataset contains 153 unique target words, with 10 (nouns and adjectives) or 20 (verbs) sample sentences being selected from the German Wikipedia for annotation. About half of this data (1040 sentences) is released as training data and is available at the time of writing.

In this work, we apply the current state of the art for English lexical substitution to this German dataset. In Section 2 we briefly cover the related work in lexical substitution. Section 3 discusses German lexical resources for obtaining substitution candidates and evaluates their suitability to the task at hand. In Section 4 we describe the final system and report on the results in Section 5.

## 2 Related Work

### 2.1 Unsupervised systems

Unsupervised approaches to the lexical substitution task typically use a contextualized word instance representation and rank substitute candidates according to their similarity to this representation. Early methods employed syntactic vector space models (Erk and Padó, 2008; Thater et al., 2011) or a clustering of instance representations (Erk and Padó, 2010). Later approaches have explored various other models, including probabilistic graphical models (Moon and Erk, 2013), LDA topic models (O Séaghdha and Korhonen, 2014), graph centrality (Sinha and Mihalcea, 2011), and distributional models (Melamud et al., 2015a).

[1] *GermEval-2015: LexSub*: https://sites.google.com/site/germeval2015/

A recent line of research takes advantage of word embeddings, which are low-dimensional continuous vector representations popularized by the skip-gram model (Mikolov et al., 2013). A simple but effective embedding-based model for lexical substitution is proposed by Melamud et al. (2015b): They decompose the semantic similarity between a target and a substitute word into a second-order target-to-target similarity based on their similarity in the embedding space, and a first-order target-to-context similarity. For this, they consider the learned context embeddings (which are usually discarded after training a Skip-gram model) and compute a substitute-to-context similarity. They achieve state-of-the-art results by just considering a (balanced) geometric mean of these two components.

### 2.2 Supervised systems

Supervised systems can be divided into *per-word* systems, which are trained on target instances per lexeme, and *all-words* systems, which aim to generalize over all lexical items. It could be shown that per-word supervised systems perform very well (with a precision > 0.8 on SemEval-2007 data) given a sufficient amount of training data for the target lexemes (Biemann, 2013). The downside of this approach is the inability to scale to unseen targets. A successful remedy to this is proposed by Szarvas et al. (2013) by the use of *delexicalized features*. The features extracted from the training data is generalized in such a way that it can generalize across lexical items beyond the training set. In this work, we build upon this framework and apply delexicalized features to German lexical substitution.

## 3 Candidate set evaluation

The lexical substitution task generally relies on lexical semantic resources to obtain a set of substitution candidates for a given lexeme. Most prevalently, WordNet (Fellbaum, 1998) is chosen as a standard resource for the English version of this task. Given multiple resources, a supervised combination of all resources was found to lead to the best results (Sinha and Mihalcea, 2009).

*GermaNet* (Hamp and Feldweg, 1997) can be considered an out-of-the-box replacement for *WordNet*. It groups lexical units into *synsets* and denotes semantic relations between these synsets. To obtain a candidate set from *GermaNet*, clearly synonyms of the substitute target should be considered (all lexemes sharing a common synset). It is further reasonable to consider both hyponyms and hypernyms of the target, as well as the transitive hull (*Transporter* → *Automobil* → *Fahrzeug* → ..) of these relations. Although higher level nodes of the *GermaNet* taxonomy include highly abstract terminology (.. → *Artefakt* → *Objekt* → *Entität*), no effort was done to exclude these terms from the candidate set. For this candidate extraction stage, no sense disambiguation of target words is performed and all senses of a given target lemma are aggregated into the candidate list.

| **candidate set** | ***R*** | ***P*** |
|---:|---|---|
| GermaNet syn | 0.05 | 0.15 |
| GermaNet syn + hy | 0.14 | 0.15 |
| GermaNet syn + hy + ho | 0.17 | 0.09 |
| GermaNet all (transitive) | 0.20 | 0.04 |
| Wiktionary | 0.17 | 0.14 |
| Woxikon | 0.44 | 0.08 |
| Duden | 0.34 | 0.15 |
| Wortschatz | 0.40 | 0.07 |
| all lexical resources | 0.61 | 0.04 |
| DT (top 200 similar) | 0.46 | 0.01 |
| DT + lexical resources | 0.71 | 0.02 |

Table 1: Candidate set evaluation on GermEval training data. The abbreviations *syn*, *hy*, and *ho* specify synonyms, direct hypernyms and direct hyponyms respectively, whereas *all* refers to pairs with an arbitrary semantic relation between them

We use UBY (Gurevych et al., 2012) to access *GermaNet* (version 9.0) and *Wiktionary*[2]. Additionally we crawl lexical resources available on the web: *Woxikon*[3], *Duden*[4] and *Leipzig Wortschatz*[5]. From these websites we scrape all listed synonyms, and in case of *Leipzig Wortschatz* all their semantic relations such as *referenced-by*, *compared-to*, and *Dornseiff-Bedeutungsgruppen* (Dornseiff, 1959).

In order to evaluate the suitability of each of these resources to the GermEval task, we construct a binary test set: each substitute pair which is present at least once in the gold data is considered a “good” expansion, whereas substitute pairs not present in the gold data are considered “bad”. For each resource, we consider the recall and precision of “good” expansion pairs, as shown in Table 1. As we perform ranking on the given candidate sets,

[2] `https://www.wiktionary.org/`

[3] `http://www.woxikon.com/`

[4] `http://www.duden.de/`

[5] `http://wortschatz.uni-leipzig.de/`

we are mostly interested in the recall, as it constitutes an upper bound for the final system. We also perform a preliminary error analysis of available substitution candidates: while all target words, and 85% of their substitutes were found in *GermaNet*, only for 20% of the GermEval pairs a semantic relation existed between these pairs. This indicates that the main problem with obtaining substitution candidates from a semantic resource is not necessarily its lexical coverage, but missing semantic relations between substitution pairs.

As an alternative to using a lexical semantic resource, fully knowledge-free approaches to lexical substitution have been proposed by the use of a distributional thesaurus (DT) (Biemann and Riedl, 2013). Although we do not follow this direction in-depth in the scope of this work, we observe that candidates obtained from a DT already yielded a better coverage than any lexical resource ($R = 0.4$) when pruned to the 200 most similar words. In line with the findings in Biemann and Riedl (2013) these candidates do not yield competitive performance within our system when compared to knowledge-based substitutes and we leave this direction open as future work.

## 4 System setup

Our system is roughly equivalent to *LexSub*[6] (Szarvas et al., 2013), although a reimplementation was used to obtain the experimental results. We follow their approach of ranking a given set of candidates based on a small set of training examples using delexicalized features. The ranking problem is cast into a binary classification task by labeling all lexical substitutions with their presence in the gold data. Hence, all substitutes which occur at least once as a gold item for a given instance are used as positive examples, whereas the remaining substitutes based on the candidate set are negative examples. We use a Maximum Entropy classifier[7] and obtain a ranking score based on the posterior probability of the positive label.

As a pre-processing step we only apply tokenization and part-of-speech tagging. We obtain the lemmatized target words directly from the gold data and have no further need to lemmatize all lexical items within the sentence, nor for syntactic parsing.

### 4.1 Features

We use most features from *LexSub*, and therefore do not cover in detail here those which can be easily adapted.

**Frequency features** A language model is used to obtain *frequency ratio* features, where an *n*-gram sliding window around a target $t$ is used to generate a set of features $\frac{freq(c_l,s,c_r)}{freq(c_l,t,c_r)}$, where $c_l$ and $c_r$ is the left and right context of $t$. We also include the different normalization variants of this feature as described in Szarvas et al. (2013), and the conjunctive phrase ratio based on the conjunctions {"*und*","*oder*",","}. For obtaining frequency counts, we evaluated 5-gram counts from *web1t* (Brants and Franz, 2009) and *German Web Counts* (Biemann et al., 2013), which both yielded nearly equivalent results.

**DT features** We create a DT from a German news corpus of 70 million sentences (Biemann et al., 2007) and obtain first-order context-features, as well as a second-order word-to-word similarity measure as described in Biemann and Riedl (2013): We prune the data, keeping only the 1000 most salient features according to a log-likelihood test (Dunning, 1993) and obtain a ranked list of 200 similar terms for each word in the corpus, based on the overlap in these context features. In particular we use as context features tuples of left and right neighbors (*de_70M_trigram*) as well as dependency features obtained using the Mate-tools[8] parser (*de_70M_mate*) to construct two distinct DTs[9].

We define delexicalized features based on the overlap in the top $k$ shared similar words ($k$ = 1, 5, 10, 20, 50, 100, 200) and top $k$ shared salient features ($k$ = 1, 5, 10, 20, 50, 100, 1000) and directly use the similarity measure between target and substitute as a feature. Lastly, we define a feature based on the accumulated LL significance measures of DT context features occurring in the sentential context. Their computation is equivalent to *coocurence* features which are explained next.

**Cooccurence features** We obtained word co-occurrence counts as described in Quasthoff et al. (2006) and define the following features: For a given sentence regarded as a

[6] Original *LexSub* system: `https://sourceforge.net/projects/lexsub/`

[7] We use the *MaxEnt* implementation of Mallet: `http://mallet.cs.umass.edu/`

[8] `https://code.google.com/p/mate-tools/`

[9] The DTs are available at `https://sourceforge.net/projects/jobimtext/files/data/models/`

bag-of-words $S$, target word $t$ and candidate set $C$, we consider the set of context words $W = S \setminus \{t\}$. For each substitute $s \in C$ we then compute the feature

$$\frac{\sum_{w \in W} LL(s,w)}{\sum_{s' \in C, w \in W} LL(s',w)}$$

where $LL$ is the log-likelihood measure of coocurence. We also compute a simple overlap version $|Co_s \cap W| / |W|$, where $Co_s$ denotes the set of words co-occurring with the substitute $s$.

**Embedding features** We roughly follow Melamud et al. (2015b) to define features in a word embedding space. To obtain German word embeddings we run the *word2vec*[10] toolkit to obtain a *CBOW* model with default parameters (200 dimensions, window-size of 8) on our German news corpus. Based on this embedding, we define two features: A second-order similarity measure between target and substitute based on cosine distance in the embedding space, as well as a very simple contextualized first-order target-to-context similarity measure. In contrast to Melamud et al. (2015b), we do not use the internal context embeddings to compute a similarity to the syntactic dependents of a target, and our embeddings are not syntax-based (Levy and Goldberg, 2014). Instead, we directly compute the similarities between a target word and a given set of context words in the embedding space, based on an $n$-gram sliding sliding window around the target. This is analogous to the delexicalized $n$-gram frequency features: For a given $n$-gram window around a target word $t$, with the context words $c_1 \ldots c_k, t, c_{k+2} \ldots c_n$, we compute for each substitute $s$ the difference in similarity to the context words with respect to the target $t$:

$$\sum_{i \leq n} |\, cos(v_s, v_{c_i}) - cos(v_t, v_{c_i}) \,|$$

where $v_x$ denotes the embedding of $x$. This is motivated by the assumption that a substitute word should behave in the same way to each context word, as the original target $t$.

**Semantic resource features** As illustrated in Section 3 we make use of various semantic relation labels from multiple semantic resources. For each lexical resource, we obtain a set of labels for a given pair of lexemes and prefix it with the name of the resource. For *GermaNet* relations, we additionally encode the length of the transitive chain, denoting an $n^{th}$-level hyponymy/hypernymy relation. For instance, the semantic relation labels for the pair $(\textit{wünschen.v}, \textit{postulieren.v})$ are {`gn_hypernym_2`, `Wortschatz_synonym`}.

| dataset | *mean* (1 − dice coefficient) | | | |
|---|---|---|---|---|
| | noun | verb | adj | all |
| SemEval-2007 | 0.750 | 0.830 | 0.755 | **0.760** |
| GermEval-2015 | 0.594 | 0.667 | 0.604 | **0.645** |

Table 2: Degree of variation within lexical substitution gold answers

Some features were discarded from the original *LexSub* system, as they could not directly be ported to German resources, or they did not prove useful. This includes the number of senses of target and substitute within *GermaNet*, the path between target and substitute within *GermaNet*, and binary features for their respective synset IDs.

## 5 Experimental results

As a preface to our evaluation, we comment briefly on the GermEval data. Upon inspection we noted that very few target lexemes in fact exhibit an ambiguous behavior. Most training instances refer to the same (or a close) meaning of a given target word, resulting in a low variance in gold answers between multiple instances of the same lexeme. We quantify this statement by calculating the mean dice coefficient between all pairwise sets of gold answers for a given lexeme. In Table 2 we compare these results to the SemEval-2007 data and observe a much lower degree of variation. A consequence of this is that a lexical substitution system based on GermEval data is less reliant on sentential context, and is primarily influenced by good prior expansions for a given word. In fact, we report a high performance on the ranking-only task (GAP=84.16% with candidate oracles), which is in line with our expectations.

**System evaluation** For evaluating the final system we perform a 10-fold cross-validation (splitting is based on target lexeme level) on the training data and report on the measures $\text{P}_{best}$, $\text{P}_{oot}$, GAP as provided by the official *GermEval* scoring tool. We disregard any multiword expressions in the gold data, as none of our candidate sets included any viable multiword expression present in the training set, and their inclusion negatively impacted results. We considered various lexical resources as potential candidate sets filtered to only single-word

[10] `https://code.google.com/p/word2vec/`

| candidate set | $P_{best}$ | GAP | P@1 |
|---|---|---|---|
| GermaNet | **15.04** | **19.12** | **55.77** |
| Wortschatz | 12.26 | 14.84 | 19.39 |
| Duden | 6.41 | 12.25 | 24.74 |
| Woxikon | 4.09 | 10.25 | 22.44 |
| Wiktionary | 3.22 | 7.50 | 22.53 |
| *candidate oracle* | 28.06 | 84.16 | (100) |

Table 3: Evaluation of the final system using different lexical resources as substitution candidates

|  | GN candidates |  |  |
|---|---|---|---|
|  | $P_{best}$ | $P_{oot}$ | GAP |
| w/o frequency feat. | 13.43 | 24.44 | 16.80 |
| w/o DT feat. | 14.77 | **24.67** | 17.59 |
| w/o sem. relation feat. | 12.26 | 23.22 | 14.84 |
| w/o embedding feat. | 14.26 | 24.64 | 17.73 |
| w/o POS feat. | 13.18 | 24.60 | 16.95 |
| full system (train-cv) | **15.04** | 24.35 | **19.12** |
| full system (testset) | 11.20 | 19.49 | 15.96 |

Table 4: Final system results and feature ablation using 10-fold cross-validation on the training set and final results

expressions. Table 3 shows the output of the full system, restricted to candidates of each resource. Despite their promising coverage of gold items in the training data (see Table 1), all lexical resources perform notably worse than *GermaNet*. This may be due to the nature of these resources: Whereas the candidate set from *GermaNet* is very accurate in enforcing the denoted semantic relationship. e.g. in case of synonymy, the other resources contain a much broader spectrum of terms that are considered "synonymous". Furthermore, the false positives in the *GermaNet* candidate set contain very obscure terms from upper levels in the ontology (*Artefakt*, *Objekt*, ..) which are easily downranked - the ranking of e.g. *Duden* candidates appears to be more difficult, as they contain mostly words which are in fact suitable in the given context. We also compare the performance to a *candidate oracle*, which serves as an upper bound for candidate sets as well as a general evaluation for the ranking-only task. Despite the bad performance as candidate sets, we find that extracting the *semantic relations* from all of these lexical resources as a feature could still notably improve the final system performance.

We further perform feature ablation test for the full system using *GermaNet* candidates as shown in Table 4. Although some features seem to exhibit redundancy (e.g. DT features and semantic relation features) all features yield a significant relative gain. It can be seen that the addition of semantic relation features yielded a relative improvement of nearly 23% for $P_{best}$, indicating that this is a strong feature for German lexical substitution. Final performance on the testset (see Table 4) is significantly worse ($P_{best} = 11.20$ compared to $P_{best} = 15.04$ on the training set with cross-validation). The reason for this is partly that candidates obtained from GermaNet have less coverage of the test data, and the test data containing more (non-covered) multiword expressions. However, when exchanging the datasets, a reasonable performance is obtained ($P_{best} = 14.68$) indicating that the issue is not related to a discrepancy between the datasets. Instead, the testset may contain generally harder instances.

## 6 Conclusion and Future Work

In this work we have successfully applied state of the art methods to German lexical substitution. We find that approaches applicable to the English version of this task can be readily adapted to German. We experimented with various lexical resources which can be used in place of their conventional English counterparts, and observe that *GermaNet* is a high quality resource which has however slight shortcomings in terms of coverage. We observe that in particular in the case of *GermaNet*, obtaining lexical substitution candidates based on the semantic relations *synonymy*, *hyponymy* and *hypernymy* is not sufficient for matching the substitutes provided by human annotators. Extracting semantic relations from other lexical resources notably improved system performance. While this is a delexicalized feature that is sufficient to generalize across all German lexical items, it is very language-dependent. In future work, we plan to overcome this dependency by generalizing features even more and experiment with delexicalized features in a multilingual setting. Additionally, we aim for a completely knowledge-free approach, obtaining substitution candidates from large background corpora.

## Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1."

## References

Chris Biemann and Martin Riedl. 2013. Text: Now in 2D! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig Corpora Collection: monolingual corpora of standard size. In *Proceedings of Corpus Linguistics*, Birmingham, UK.

Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski, and Torsten Zesch. 2013. Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics*, 28(2):23–60.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Thorsten Brants and Alex Franz. 2009. Web 1T 5-gram, 10 European languages version 1. *Linguistic Data Consortium*.

Kostadin Cholakov, Chris Biemann, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Lexical Substitution Dataset for German. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Franz Dornseiff. 1959. *Der deutsche Wortschatz nach Sachgruppen*. Walter de Gruyter.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Waikiki, Honolulu.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 conference short papers*, pages 92–97, Uppsala, Sweden.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France.

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

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308, Baltimore, USA.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53, Prague, Czech Rep.

Oren Melamud, Ido Dagan, and Jacob Goldberger. 2015a. Modeling Word Meaning in Context with Substitute Vectors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, Denver, USA.

Oren Melamud, Omer Levy, Ido Dagan, and Israel Ramat-Gan. 2015b. A Simple Word Embedding Model for Lexical Substitution. *VSM Workshop*. Denver, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, Harrah's Lake Tahoe, USA.

Taesun Moon and Katrin Erk. 2013. An inference-based model of word meaning in context as a paraphrase distribution. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):42.

Diarmuid O Séaghdha and Anna Korhonen. 2014. Probabilistic Distributional Semantics with Latent Variable Models. *Computational Linguistics*, 40(3):587–631.

Uwe Quasthoff, Matthias Richter, and Christian Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genova, Italy.

Ravi Sinha and Rada Mihalcea. 2009. Combining lexical resources for contextual synonym expansion. In *Proceedings of the Conference in Recent Advances in Natural Language Processing*, pages 404–410, Borovets, Bulgaria.

Ravi Som Sinha and Rada Flavia Mihalcea. 2011. Using Centrality Algorithms on Directed Graphs for Synonym Expansion. In *FLAIRS Conference*, Palm Beach, USA.

György Szarvas, Chris Biemann, Iryna Gurevych, et al. 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*, pages 1131–1141, Atlanta, USA.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word Meaning in Context: A Simple and Effective Vector Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1134–1143, Stroudsburg, PA, USA.

# Lexical substitution using deep syntactic and semantic analysis

**Luchezar Jackov**
Institute for Bulgarian Language
Bulgarian Academy of Sciences
lucho@skycode.com

## Abstract

This paper presents an approach for lexical substitution for the GermEval 2015 shared task using the machine translation (MT) system presented by Jackov (Jackov, 2014). The system performs deep transfer using German lexicalisations of the Princeton WordNet (PWN) (Fellbaum, 1998) synsets as a part of the deep syntactic and semantic internal representation of the input text. The analysis step of the system is used to disambiguate the head word. Once it is disambiguated in terms of PWN synset id, synonym and hypernym lexemes are used as substitution candidates.

## 1 Introduction

The lexical substitution task consists of finding appropriate substitutes for a given word in a given context and ranking them by appropriateness. This task sets no restrictions on the approaches for tackling with it as it does not require a specific sense inventory.

The data for the GermEval 2015: LexSub task is described by Cholakov et al. (2014). 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 (Cholakov et al., 2014).

Half of the data has been provided by GermEval 2015 organisers in advance as a training set, while the other half was used for evaluation.

## 2 Previous and related work

GermEval 2015 is inspired by the English lexical substitution task (McCarthy and Navigli, 2009). The original aim of the task organised at SemEval 2007 was to provide a WSD evaluation where the sense inventory is not predefined, allowing for much wider range of systems to participate.

The lexical substitution task faces two main problems: one is the generation of possible substitutes and the other is their ranking. Some researchers focused only on the ranking problem while others tried to address both.

A detailed and structured overview of the related work is given by Szarvas et al. (Szarvas et al., 2013).

## 3 Proposed approach

An interesting approach for deep syntactic and semantic disambiguation was presented by Jackov as part of an MT system. The internal interpretation of the input text uses PWN synsets, which makes it easy to use it for the lexical substitution task once a PWN synset id is identified for the head word.

The proposed disambiguation approach considers the input text as a sequence of tokens. Then for each token all possible lemmas are derived. Lemma sequences of 1 or more tokens are looked up by the concept binder module in a synset lexicalisation table for PWN synsets. Each successful look-up is an assumption for a concept and constitutes an initial parsing hypothesis. The hypotheses contain assumptions about the concepts lying behind the input tokens, their syntactic roles and their dependency relations. Adjacent hypotheses are combined into new hypotheses for larger spans of the input sequence by using manually written hypothesis derivation rules. Each rule identifies, inherits and extends the syntactic and semantic assumptions of the constituting hypotheses. The rules are applied using a modified version of the Cocke–Younger–Kasami (CYK) algorithm (Cocke et al., 1970; Younger, 1967; Kasami, 1965) until all spans of the input sequence are covered. To prevent hypothesis space explosion each hypothesis is scored against a knowledge database of dependency relations and only the n-best hypotheses are kept for each span of tokens.

When the hypothesis generator finishes its work it yields a parsing hypothesis for the input sequence of tokens having the best score.

The internal representation of the input sequence within the hypothesis contains a PWN synset id for each of the concepts that form the hypothesis, including a concept for the head word. The synset id of the concept is used to derive substitution candidates for the head word.

The goal of this paper is to evaluate the system's WSD module when using it as a lexical substitution tool.

## 4 Detailed description of the MT system used for disambiguation

### 4.1 Overview

The system has been implemented in C++ and has a very compact binary data representation, approx. 120MB for 7 languages and 42 language translation directions. It has been used in offline translation applications for mobile devices, outperforming Google Offline Translator in both quality and size (the latter needs about 1.05GB of data for the same 7 languages). It has also participated successfully in the iTranslate4 project, and can be tested online at http://itranslate4.eu (the SkyCode vendor). The system consists of a lemmatizer, a concept binder, a hypothesis generator, a dependency relations scorer and a synthesis unit (Jackov, 2014).

The system implements an extensive inventory of categories and category values. A special category, the hypothesis type identifier (HTI), serves as the set of non-terminal values for the parsing rules, which are extended context-free grammar (CFG) rules used for production of hypotheses.

An elaborate description with many examples is given by Jackov (Jackov, 2014).

### 4.2 Lemmatizer

The first step of the system operation is to apply the lemmatizer module on each input token, which produces a list of all lemmas for each token along with their category values. The lemma of each lemmatization is kept as a lemma identifier, which is used later in the concept binder module. The lemmatizer is built as a simple, yet very efficient stemmer allowing definition of arbitrary paradigms, one per HTI. The original system has 144,243 lemmas for German.

### 4.3 Hypothesis generator

The second step is to apply the hypothesis generator for every span of the input sequence of tokens. The module first runs the concept binder for spans of length less than 7 tokens, and then applies parsing rules over the adjacent sub-spans of each span.

### 4.4 Concept binder

The concept binder finds the concepts (PWN synset ids) that match a span of input tokens.

It uses a database of the possible lexicalisations for each PWN synset. Each lexicalisation entry in the database consists of a list of lemma identifiers, PWN synset id, attribute restriction rules, attribute unification rules, and a list of additional attribute values. The list of additional values is used to define lexicalisation level features such as sub-categorization frames, transitiveness and aspect for verbs, etc. The original system has 229,575 synset lexicalisations for German.

The PWN synset lexicalisations for the six languages other than English have been automatically gathered from various sources and manually improved for the goal of creating a multi-language MT system.

### 4.5 Parsing rules and hypothesis generation

The core of each parsing rule is an extended CFG rule defined for the HTI feature values of the constituting hypotheses. The parsing rule extends the CFG by defining additional attribute value restrictions, agreement restrictions, attribute unification rules and parsing rule score. It also defines syntactic and semantic roles, dependency relations and propagation rules so that the higher level hypothesis resulting from the rule application unifies those of the constituting hypotheses (see Figure 1 below).

### 4.6 Dependency relations knowledge database

The database contains entries that consist of a relation identifier, two PWN synset ids and a weight value, which is normally 1 or -1.

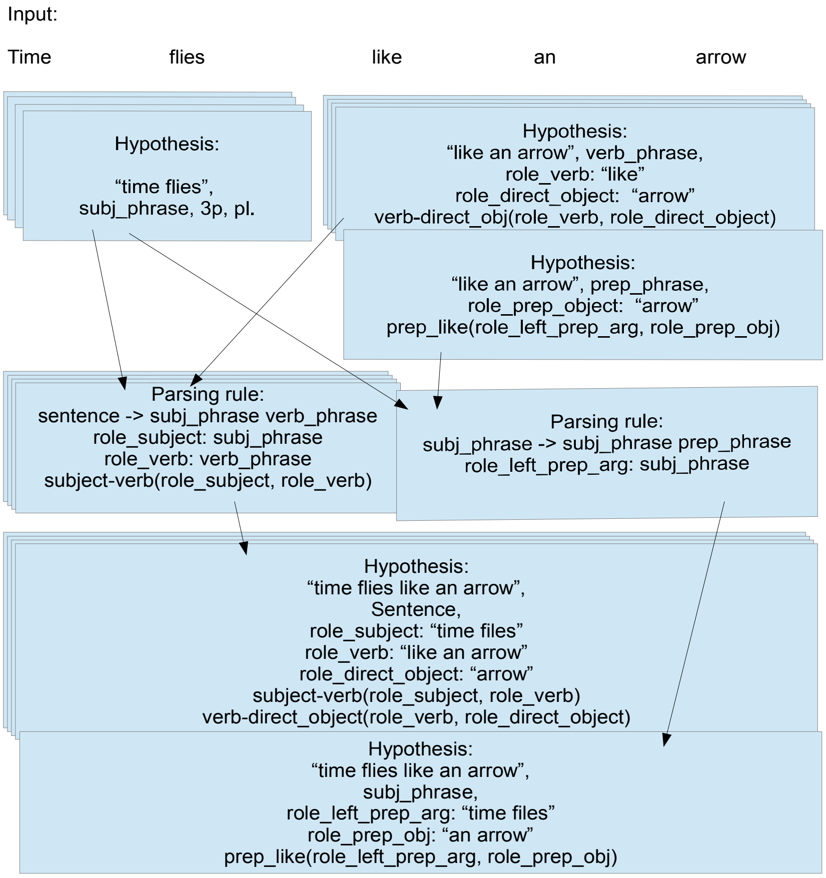
Input:
Time
flies
like
an
arrow
Hypothesis:
“time flies”,
subj_phrase, 3p, pl.
Hypothesis:
“like an arrow”, verb_phrase,
role_verb: “like”
role_direct_object: “arrow”
verb-direct_obj(role_verb, role_direct_object)
Hypothesis:
“like an arrow”, prep_phrase,
role_prep_object: “arrow”
prep_like(role_left_prep_arg, role_prep_obj)
Parsing rule:
sentence -> subj_phrase verb_phrase
role_subject: subj_phrase
role_verb: verb_phrase
subject-verb(role_subject, role_verb)
Parsing rule:
subj_phrase -> subj_phrase prep_phrase
role_left_prep_arg: subj_phrase
Hypothesis:
“time flies like an arrow”,
Sentence,
role_subject: “time files”
role_verb: “like an arrow”
role_direct_object: “arrow”
subject-verb(role_subject, role_verb)
verb-direct_object(role_verb, role_direct_object)
Hypothesis:
“time flies like an arrow”,
subj_phrase,
role_left_prep_arg: “time files”
role_prep_obj: “an arrow”
prep_like(role_left_prep_arg, role_prep_obj)

**Figure 1.** Parsing rules being applied to hypotheses yield hypotheses for broader spans. Even though the illustrated hypotheses seem unlikely for the sample input text, this may not be so for other input text (e.g., “Time travels seem an illusion”). The likeliness is evaluated as a hypothesis score by looking up the identified dependency relations in the knowledge base. Note that this figure shows just one of the possible splits and other splits are also considered, such as the semantically correct one, [S → SP VP (“time”, *subj_phrase*) (“flies like an arrow”, *verb_phrase*)]. When having good knowledge base, the latter hypothesis will receive the best score.

The database is manually populated and currently has 1,803,446 entries. As the relations are defined over PWN synset ids, they can be used for all languages for which synset lexicalisations exist.

Here are sample entries with words instead of PWN synset ids for clarity:

(poss, study, woman, 1)

(nsubj, mushroom, study, 1).

The above entries are enough for disambiguating the sentence *Women's studies mushroom.* This is an actual headline, which many humans find hard to comprehend, meaning that the studies done by women grow rapidly.

### 4.7 Hypothesis scoring

As a result each hypothesis contains a number of assumed concepts and their dependency relations and each concept is identified by its PWN synset id. The set of the relations between the concepts is scored by looking up the dependency relations knowledge base. If the look-up is successful the dependency relation score is the weight of the matching entry, otherwise the score is zero. The hypothesis score is calculated by summing the dependency relation scores and the parsing rule score.

## 5 Derivation of lexical substitutions

A parsing hypothesis having the best score is obtained as a result of applying the analysis modules described above. The hypothesis contains PWN synset ids for each concept that has been identified and each concept is covered by one or more tokens. The concept covered by the head word is used to derive synonyms and hypernyms to be used as lexical substittions.

The final list of substitutions is formed by the list of synonym lexemes followed by the list of hypernym lexemes ordered by the usage count data from the synset lexicalisations table.

## 6 Shortcomings of the approach

The proposed approach uses PWN 3.0 synset definitions which are best fit to English. There are lexical gaps between English and German that cannot be addressed properly using the current level of detail in PWN.

The described approach sorts the candidates by lexicalisation usage count, while this may not be the most appropriate metric for lexical substitution. Currently only direct synonyms and hypernyms are used, while in many cases using near synonyms may yield additional good substitutions.

## 7 Future work

The advantages of the system used for WSD can be exploited better in several ways. One way is to rank the substitution candidates by using each candidate in the input text and evaluating the text by getting the best hypothesis score.

Another direction for future work is to evaluate the substitution candidates using statistics of the dependency relations over lemmas made over large monolingual corpus, thus capturing finer differences between the word meanings within the PWN synonym sets. Jackov has mentioned about using this approach in order to improve translations in his system (Jackov, 2014).

Yet another direction is to generate the substitution candidates list using the above-mentioned statistics.

## 8 Results and observations

The following table shows the results of this system on the test data compared to the baseline systems  results provided on the GermEval site.

| System | Best | | OOT | | GAP |
|---|---|---|---|---|---|
| | P | R | P | R | |
| Jackov | 6.73 | 6.45 | *20.14* | *19.32* | 0.1126 |
| RSense | 7.40 | 7.40 | 12.53 | 12.53 | 0.0954 |
| TRS | **10.04** | **10.04** | 15.21 | 15.21 | 0.1225 |
| WSense | 7.50 | 7.50 | 20.54 | **20.54** | **0.1428** |
| Jackov/m | 13.36 | 12.86 | *33.18* | *31.92* | 0.1126 |
| RSense/m | 15.13 | 15.13 | 23.45 | 23.45 | 0.0954 |
| TRS/m | **19.82** | **19.82** | 27.99 | 27.99 | 0.1225 |
| Wsense/m | 13.46 | 13.46 | **35.55** | **35.55** | **0.1428** |

There scoring methodologies have been used: best, out-of-ten (OOT), and general average precision (GAP).

The poor results for the 'best' metric clearly show that the chosen ranking criterion is not adequate. This could be explained by the fact that the lexicalisations for German are gathered semi-automatically from unannotated corpora and the lexicalisation usage count is more often than not set to zero in the lexicalisation table.

The good OOT results show that the WSD module of the system performs reasonably well.

## 9 Conclusion

In this article we have presented the use and evaluation of a deep syntactic and semantic analysis system for the task of lexical substitution for German. The approach relies on syntactically and semantically driven dependency parsing using PWN lexicalisations for German for both disambiguation and derivation of substitution candidates. The results demonstrate that the proposed approach is a viable method for both word sense disambiguation and lexical substitution. It can be improved further in several ways, leading to supposedly better lexical selection.

## References

**Cholakov et al., 2014:** K. Cholakov, C. Biemann, J. Eckle-Kohler, and I. Gurevych. Lexical substitution dataset for German. In *Proceedings of the 9th International Conference on Language Resources and Evaluations* (LREC 2014), pages 1406–1411. May 2014.

**Cocke et al., 1970:** J. Cocke, and J. T. Schwartz. 1970. *Programming Languages and Their Compilers: Preliminary Notes.* Technical report. Courant Institute of Mathematical Sciences, New York University.

**Fellbaum, 1998:** C. Fellbaum. 1998. WordNet: An Electronic Lexical Database. MIT Press.

**Jackov, 2014:** L. Jackov. 2014. Machine translation based on WordNet and dependency relations, In *Proceedings of Computational Linguistics in Bulgaria 2014*, pages 64-72.

**Kasami, 1965:** T. Kasami. 1965. An Efficient Recognition and Syntax-analysis Algorithm for Context-free Languages. *Scientific report AFCRL-65-758.* Bedford, MA: Air Force Cambridge Research Lab.

**McCarthy and Navigli, 2009:** D. McCarthy and R. Navigli. The English lexical substitution task. *Language Resources and Evaluation* 43:139–159. 2009.

**Szarvas et al., 2013:** G. Szarvas, C. Biemann, and I. Gurevych. 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. 2013.

**Younger, 1967:** D. H. Younger. 1967. Recognition and parsing of context-free languages in time n3. *Information and Control 10 (2)*: 189-208.