Zmorge: A German Morphological Lexicon Extracted from Wiktionary

Rico Sennrich, Beat Kunz

Institute of Computational Linguistics
University of Zurich
Binzmühlestr. 14
CH-8050 Zürich
sennrich@cl.uzh.ch, beat.kunz@access.uzh.ch

Abstract

We describe a method to automatically extract a German lexicon from Wiktionary that is compatible with the finite-state morphological grammar SMOR. The main advantage of the resulting lexicon over existing lexica for SMOR is that it is open and permissively licensed. A recall-oriented evaluation shows that a morphological analyser built with our lexicon has comparable coverage compared to existing lexica, and continues to improve as Wiktionary grows. We also describe modifications to the SMOR grammar that result in a more conventional lemmatisation of words.

Keywords: German morphology, lexicon, collaborative resource construction

1. Introduction

Resources for morphological analysis should ideally be open, permissively licensed, have a wide coverage, and be regularly updated to reflect language change. Current German morphology analysers fail to meet one or several of these requirements. The most open tool, Morphisto (Zielinski and Simon, 2009), combines the SMOR grammar (Schmid et al., 2004) with an open lexicon, but the lexicon is only licensed for non-commercial use, and no scalable workflow is in place to maintain and extend it.

To address these issues, we present a tool to automatically extract a German morphological lexicon from Wiktionary. Wiktionary is open, permissively licensed, and has a a respectable size, with about 48 000 noun stems and 5500 verb stems at the time of this writing. Also, the crowd-sourced architecture of Wiktionary and its active community ensure that the lexicon will be updated to include new word forms, and reflect future changes in orthography.

The result of our extraction tool is a morphological lexicon that is compatible with the SMOR grammar, and can thus be compiled into a finite-state morphological analyser. Finite-state morphological analysers are important for processing morphologically productive language such as German, and SMOR has been used to improve NLP tasks such as parsing (Seeker et al., 2010; Sennrich et al., 2013) and statistical machine translation (Fritzinger and Fraser, 2010; Williams and Koehn, 2011).

2. Related Work

We give a short overview of German morphology analysers, and attempts to automatically build morphological lexica for German.

Two examples of closed-source morphology tools, which both require licensing, are the commercial tool GERTWOL (Haapalainen and Majorin, 1995), and Stripey Zebra (Lorenz, 1996; Schulze, 2004).

Schmid et al. (2004) present SMOR, a German finite-state morphology. The grammar itself has been made available as free software under the GPL v2, but the lexicon is closed-source, and only a compiled transducer is available for academic research.

Adolphs (2008) describes a method to acquire an inflectional lexicon for the SMOR morphology from unannotated corpora. He uses a modified version of SMOR to create inflectional hypotheses for each word form in the corpus, and then selects a hypothesis based on frequency statistics in the corpus. A problem with this approach is that not all inflectional classes can be disambiguated based on the observed word forms alone – some noun classes share the same ending, but have different genders or use the same ending for different grammatical cases. An example is *-s*, which could be a genitive or a plural marker, or both.

Zielinski and Simon (2009) have produced an open lexicon for the SMOR grammar. Their lexicon is based on several dictionaries, such as a digital edition of "Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart".[1] Morphisto is open, but under a non-commercial license, and one of our main goals is to create a lexicon with a fully permissive license. Also, we believe that a lexicon needs to be actively maintained to react to language change, and that the best way to make such an effort sustainable is through a collaborative architecture. Automatically extracting a lexicon from Wiktionary, which is an established project with an active community, is thus a natural choice to ensure that the lexicon will profit from active development in the long-term.

Recently, similar work on building a German lexicon from Wiktionary has been performed for the Alexina toolkit (Sagot and Walther, 2013).

3. Lexicon Extraction from Wiktionary

The SMOR grammar defines inflection classes that cover the morphology of the majority of German words. For instance, the inflection class "NNeut_s_e" stands for a normal

[1] `http://www.zeno.org/Adelung-1793`

case	singular	plural
nominative	das Haus	die Häuser
genitive	des Hauses	der Häuser
dative	dem Haus dem Hause	den Häusern
accusative	das Haus	die Häuser

Table 1: Inflection table of *Haus* (Engl. *house*).

```
Hauses<NNeut_0_x><+NN><Neut><Nom><Sg>
Hause<NMasc_s_s><+NN><Masc><Nom><Pl>
Haus<NMasc_es_$e><+NN><Masc><Gen><Sg>
Haus<NMasc_es_$er><+NN><Masc><Gen><Sg>
Haus<NMasc_es_er><+NN><Masc><Gen><Sg>
Haus<NNeut_es_$er><+NN><Neut><Gen><Sg>
```

Figure 1: Initial inflection hypotheses for *Hauses* (excerpt).

noun with neutral gender, -s genitive marker and -e plural marker. An example for this class is *Spiel* (Engl. game).[2]
Our aim is to create a lexicon that is compatible with the SMOR grammar by extracting the relevant information from Wiktionary, and automatically finding the correct inflection class for each entry. Some information can be extracted with simple pattern matching, e.g. part-of-speech and gender information. The main challenge is an automatic mapping of the inflection tables in Wiktionary to its corresponding inflection class in SMOR.

3.1. Inflection Prediction

Wiktionary provides us with an inflection table as shown in table 1, and our aim is to automatically find the SMOR inflection class that matches the word forms in the inflection table. We base our approach on SLES, a module for generating SMOR inflection class hypotheses from word forms (Adolphs, 2008).
Adolphs uses SLES for automatic lexicon acquisition from a corpus by creating inflectional hypotheses for each word form in the corpus, and then selecting the inflection class for the lexicon based on frequency statistics. A problem with this approach is that not all inflectional classes can be disambiguated based on raw text alone: we may not be able to determine the gender of a lemma, or whether -s is used as a genitive marker or plural marker, or both.
Performing inflection prediction with data extracted from Wiktionary rather than raw text has several advantages. We avoid all uncertainty as to whether two word forms belong to the same lemma or not, and we know the grammatical category of each word form in the inflection table.
Our approach for hypothesis selection is as follows: For each word form in an inflection table, we generate a list of inflection class hypotheses with SLES, and filter this list through information gathered from Wiktionary, e.g. part-of-speech, stem, gender or case.

[2]There are special inflection classes that allow the handling of irregular words, although this requires multiple entries for the same lemma. We adopted and extended the manually-compiled list of irregular forms by (Adolphs, 2008).

For the word *Hauses* (genitive singular of the neuter noun *Haus*), SLES returns 120 hypotheses, some of which are shown in figure 1. Note that SLES proposes inflection classes without any lexical knowledge; it thus also suggests that *Hauses* could be the nominative plural of the masculine noun *Hause*, with -s as both genitive and plural marker. We can discard this hypothesis based on structural information from Wiktionary. Filtering out hypotheses with the wrong stem, gender, case or number reduces the number of hypotheses for the genitive singular to four, shown in figure 2. In a second step, the hypotheses for all word forms in the inflection table are intersected, which ideally gives us the final inflection class – in the example of *Haus*, "NNeut_es_$er" (-es as genitive marker; umlaut and -er as plural marker).
If the resulting set of predicted inflectional classes is empty, this is typically due to the word having an irregular inflection that SMOR cannot encode in a single inflection class. We rely on manual entries for irregular verbs; for nouns, we try to predict a singular inflection and a plural inflection independently, and add them to the lexicon if both are unambiguous. Other, less frequent causes for a failure to predict an inflection class are typos in the inflection table or other noise such as mark-up information, most of which we remove in preprocessing.
If an inflection table in Wiktionary is incomplete, SMOR may predict multiple inflection classes. This commonly happens for singular nouns, plural nouns and adjectives without comparative form, to which we assign special inflection classes.
A complicating factor is dealing with Wiktionary pages that need to be mapped to multiple entries in the lexicon. Homonyms such as *See*, which can be a feminine noun (Engl. *sea*) or a masculine noun (Engl. *lake*), are listed under different headings on the same page. If a word has multiple valid inflectional paradigms, there may be additional columns in the inflection table, for instance for *Lexikon* (Engl. *lexicon*), which has the plural stems *Lexiken* and *Lexika*. We identify such cases in preprocessing and split the Wiktionary page into multiple entries for which we predict the inflection class separately.
Preprocessing is also required to identify inflectional variants, which are sometimes given in brackets, sometimes separated by commas or linebreaks. If inflection prediction results in multiple variants, we create multiple entries in the lexicon. An example is *Islam*, which may have the genitive marker -s or zero.
Note that the whole process is fully automated, so that we can easily update the lexicon as Wiktionary grows. The number of irregular words is relatively small, and we keep the manually created entries in separate files so that they can be merged into new versions of the lexicon.

4. Lexicon Evaluation

We perform a recall-oriented evaluation of our lexicon by comparing it to two other lexica that both use the SMOR grammar: Morphisto (Zielinski and Simon, 2009) and the original SMOR lexicon (Schmid et al., 2004). We call our lexicon Zmorge, which stands for Zurich Morphological

```
nom sg (Haus):    NNeut-s/$sser, NNeut-s/sse, NNeut_0_x, NNeut_es_$er,
                  NNeut_es_e, NNeut_es_en, NNeut_es_er, NNeut_s_e
gen sg (Hauses):  NNeut_es_$er, NNeut_es_e, NNeut_es_en, NNeut_es_er
dat sg (Haus):    NNeut-s/$sser, NNeut-s/sse, NNeut_0_x, NNeut_es_$er,
                  NNeut_es_e, NNeut_es_en, NNeut_es_er, NNeut_s_e
acc sg (Haus):    NNeut-s/$sser, NNeut-s/sse, NNeut_0_x, NNeut_es_$er
                  NNeut_es_e, NNeut_es_en, NNeut_es_er, NNeut_s_e
nom pl (Häuser):  NNeut_es_$er
gen pl (Häuser):  NNeut_es_$er
dat pl (Häuser):  NNeut_es_$er
acc pl (Häuser):  NNeut_es_$er
```

Figure 2: Filtered inflection hypotheses for all word forms of *Haus*. In bold: analysis resulting from intersecting the hypothesis sets of all word forms.

lexicon	NN	NE	V	ADJ	total
Morphisto	86.5	8.5	89.0	54.7	69.0
Stuttgart lexicon	90.5	29.5	97.4	59.4	76.3
Zmorge	88.6	18.4	88.6	57.0	72.1

Table 2: Evaluation of SMOR with different lexica. Percentage of word types in TüBa-D/Z which are correctly analysed.

Lexicon for German, and also means "breakfast" in Swiss German.

As data set, we use the manually annotated TüBa-D/Z treebank version 7 (Telljohann et al., 2004), looking only at verbs, adjectives, normal nouns and proper nouns, which are the morphologically complex cases. We evaluate the systems on the type-level, where two tokens are treated as the same type if they share word form, part-of-speech and morphological analysis. This gives us approximately 134 000 word types. An analysis is considered correct if at least one of the returned hypotheses has the correct lemma, part-of-speech and morphological features.

Table 2 shows the percentage of correct analyses for normal nouns (NN), proper nouns (NE), verbs (V), adjectives (ADJ), and in total. Generally, the lexical coverage of our system is higher than that of Morphisto, but lower than that of the original SMOR lexicon. We manually inspected unanalysed word forms to identify the main causes of failure. The main reason why no analysis could be found with our system was that the relevant entry was simply not in Wiktionary.[3] This is most conspicuous for verbs, where even some common verbs such as *beschuldigen* (Engl. *accuse*) are missing, and proper nouns. The German Wiktionary is still growing rapidly, so we are confident that this gap will narrow in the future.

There are a number of entries for which the lexical extraction fails because of an irregular or foreign-language inflection which is not supported by SMOR, e.g. *Appendix, -izes*. For other entries, no inflection class is found because of missing information, typos or idiosyncracies in the layout of the Wiktionary page. The problem of missing information is most prevalent for abbreviations, for which even part-of-speech information may be missing.

All three systems have a relatively high number of errors for adjectives. This can be explained by how the SMOR grammar handles derivational morphology, which makes the mapping between the lemmas in TüBa and the SMOR results difficult. For instance, for German adjectives that are derived from verbs, e.g. *gesucht* (Engl. *sought-after*), TüBa gives the adjective form *gesucht* as lemma, whereas SMOR returns the verb form *suchen*. Since this is a problem of the SMOR grammar and not the lexica, all three systems that we compare are equally affected.

5. Improving the Lemmatisation of SMOR

If SMOR analyses a word form derivationally, the analysis string shows the associated base forms, but does not correspond to what we conventionally consider the lemma of the full word, i.e. the nominative singular form for nouns, the infinitive form for verbs, and the adverbial form for adjectives. For instance, *Ermittlungen* (Engl. *investigations*) is analysed as follows by SMOR:

```
> Ermittlungen
ermitteln<V>ung<SUFF><+NN><Fem><Acc><Pl>
ermitteln<V>ung<SUFF><+NN><Fem><Dat><Pl>
ermitteln<V>ung<SUFF><+NN><Fem><Gen><Pl>
ermitteln<V>ung<SUFF><+NN><Fem><Nom><Pl>
```

The analysis shows the derivation of the word (from the verb *ermitteln* (Engl. *investigate*)), but we cannot easily extract the conventional lemma, i.e. the nominative singular form *Ermittlung*, without linguistic knowledge. In our evaluation in the previous section, we perform a heuristic mapping to a pseudo-lemma by selecting the last morpheme in the analysis string, and concatenating it with the unnormalized stem. We separate the stem which we want to retain, and the ending which we substitute with the normalized form, through a longest common subsequence match between the original word form and the last morpheme in the SMOR analysis. In the example above, the last morpheme in the analysis is *ung*, which means that our pseudo-lemma is the concatenation of *Ermittl* and *ung*, which corresponds to the correct lemma *Ermittlung*.

This heuristic is imperfect, and is only applied for nouns. Thus, all systems in our initial evaluation fail to lemmatise

[3] We performed our experiments with an XML dump of the German Wiktionary from February 2014.

system	NN	NE	V	ADJ	total
direct mapping	72.8	18.4	88.6	57.0	63.3
heuristic mapping	88.6	18.4	88.6	57.0	72.1
modified SMOR	90.9	18.4	88.5	85.2	78.3

Table 3: Evaluation of lemmatisation variants. Percentage of word types in TüBa-D/Z which are correctly analysed.

system	NN	NE	V	ADJ	total
Wiktionary 15/11/2012	89.5	17.3	86.8	83.8	77.0
Wiktionary 24/02/2014	90.9	18.4	88.5	85.2	78.3
+ list of regular verbs	91.3	18.4	92.6	86.5	79.3

Table 4: Evaluation of morphological analysers based on different Wiktionary versions. Percentage of word types in TüBa-D/Z which are correctly analysed.

deverbal adjectives, among others. As a more systematic solution, we modified the SMOR grammar to return what we consider the desirable lemma, i.e. the infinitive form for verbs, the positive adverbial form for adjectives, and the nominative singular form for nouns.

We do this by composing two transducers: the original transducer generated by SMOR, and a transducer that is derived from the original one and maps the analysis form *ermitteln<V>ung<SUFF><+NN>* to the desired lemma *Ermittlung<+NN>*. We obtain this derived transducer by filtering the original transducer to only contain base forms (nominative singular / infinitive / positive forms), stripping the grammatical features from the analysis side, and inverting the transducer.

For applications such as compound splitting, morpheme boundaries in the analysis string are valuable information. We modified the grammar so that morpheme boundaries are retained in the transducer; for *Ermittlungen*, our modified transducer yields the lemma *Ermittl<~>ung*, with separate morpheme boundary markers for German linking morphemes (*Fugenelemente*), inflectional boundaries, and compound boundaries.

Table 3 shows evaluation results for three methods of mapping SMOR analyses to lemmas:

- direct mapping: remove any categorial information like <*V*> or <~>, and consider the remainder to be the lemma.
- heuristic mapping based on longest common subsequence matching: this method was used in the evaluation in the previous section.
- the modified SMOR grammar.

The results are shown in table 3. We can see that the modified SMOR grammar is slightly better than a heuristic mapping for nouns, and yields a markedly better recall for adjectives (by almost 30 percentage points). In total, recall improves by 6 percentage points. We found some inconsistencies in the TüBa gold annotation which account for the majority of remaining recall errors for adjectives. Specifically, SMOR always maps adjectives in the comparative or superlative degree to the corresponding base form in the positive degree. In TüBa, the lemmatisation of adjectives is inconsistent, with adjectives in the comparative degree sometimes mapped to a lemma in the positive degree, sometimes not.

6. Tracking the Growth of Wiktionary

While we cannot accurately predict the future development of Wiktionary, comparing recent versions indicates that development is still very active, and that repeating the extraction in the future will allow us to profit from this development. Table 4 compares morphological analysers that were created with the same extraction script and grammar, but two versions of Wiktionary: one from November 2012, one from February 2014. In a period of 15 months, additions and corrections to Wiktionary led to an improvement of 1.3 percentage points in our morphology evaluation. The growth of Wiktionary is also reflected in our lexicon: the number of entries that were extracted grew from 58 000 to 68 000.

We also found that Wiktionary users have compiled a list of regular verbs, not all of which have their own page yet. Adding this list to the lexicon boosts the performance of our morphological analyser by a further 4.1 percentage points for verbs, and 1 percentage point in total. This can serve as an outlook as to how performance will increase as these verbs receive their own page on Wiktionary, but for practical purposes, we can already include them in our lexicon, since we can skip inflection class prediction for regular verbs.

7. Conclusion

We describe an automatic method to extract a morphological lexicon from the German version of Wiktionary that can be used with the SMOR grammar to build a finite-state morphological analyser for German. Our evaluation results show that the coverage of our lexicon is already better than that of the Morphisto lexicon, but still smaller than that of the original Stuttgart lexicon. We also present modifications to the SMOR grammar that implement a different notion of lemmatisation, i.e. returning the infinitive form for verbs, the positive (adverbial) form for adjectives, and the nominative singular form for nouns, rather than returning a derivational analysis.

The main advantage of our lexicon is that it falls under the same license as Wiktionary (CC BY-SA 3.0) and is thus more permissive than both Morphisto and the Stuttgart lexicon. Also, the fact that the extraction of the lexicon is fully automated means that we can benefit from future improvements in Wiktionary, both in terms of better coverage and reacting to language change (such as spelling changes or word formation). Pre-built lexica and transducers, and links to the source code of both the extraction script and the modified SMOR grammar, are available at `http://kitt.ifi.uzh.ch/kitt/zmorge/`.

8. References

Peter Adolphs. 2008. Acquiring a Poor Man's Inflectional Lexicon for German. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco.

Fabienne Fritzinger and Alexander Fraser. 2010. How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 224–234, Uppsala, Sweden. Association for Computational Linguistics.

Mariikka Haapalainen and Ari Majorin. 1995. GERTWOL und Morphologische Disambiguierung für das Deutsche. In *Proceedings of the 10th Nordic Conference of Computational Linguistics*, Helsinki.

Oliver Lorenz. 1996. Automatische Wortformerkennung für das Deutsche im Rahmen von MALAGA. Master's thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Benoît Sagot and Géraldine Walther. 2013. Implementing a Formal Model of Inflectional Morphology. In Cerstin Mahlow and Michael Piotrowski, editors, *Systems and Frameworks for Computational Morphology*, volume 380 of *Communications in Computer and Information Science*, pages 115–134. Springer Berlin Heidelberg.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1263–1266.

Markus Schulze. 2004. *Ein sprachunabhängiger Ansatz zur Entwicklung deklarativer, robuster LA-Grammatiken mit einer exemplarischen Anwendung auf das Deutsche und das Englische*. Ph.D. thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Wolfgang Seeker, Bernd Bohnet, Lilja Øvrelid, and Jonas Kuhn. 2010. Informed ways of improving data-driven dependency parsing for German. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1122–1130, Beijing, China. Association for Computational Linguistics.

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2013*, pages 601–609, Hissar, Bulgaria.

Heike Telljohann, Erhard W. Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 2229–2235, Lisbon, Portugal.

Philip Williams and Philipp Koehn. 2011. Agreement Constraints for Statistical Machine Translation into German. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 217–226. Association for Computational Linguistics.

Andrea Zielinski and Christian Simon. 2009. Morphisto – An Open Source Morphological Analyzer for German. In *Proceedings of the 2009 conference on Finite-State Methods and Natural Language Processing: Postproceedings of the 7th International Workshop FSMNLP 2008*, pages 224–231, Amsterdam. IOS Press.