Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4653–4657
Marseille, 11–16 May 2020

MucLex: A German Lexicon for Surface Realisation

Kira Klimt[1], **Daniel Braun**[1], **Daniela Schneider**[2], **Florian Matthes**[1]
[1]Technical University of Munich, [2]Allianz SE
Munich, Germany
{kira.klimt, daniel.braun, matthes}@tum.de
daniela.schneider1@allianz.de

Abstract

Language resources for languages other than English are often scarce. Rule-based surface realisers need elaborate lexica in order to be able to generate correct language, especially in languages like German, which include many irregular word forms. In this paper, we present *MucLex*, a German lexicon for the Natural Language Generation task of surface realisation, based on the crowd-sourced online lexicon Wiktionary. MucLex contains more than 100,000 lemmata and more than 670,000 different word forms in a well-structured XML file and is available under the Creative Commons BY-SA 3.0 license.

Keywords: Lexicon, Surface Realisation, Natural Language Generation

1. Introduction

A rule-based surface realiser can only be as good as the lexicon it is based on. This is especially the case for languages like German, which have a relatively high grade of irregular inflection, compared to e.g. English.

The German language is characterised by various irregular word inflection forms. Adjectives, nouns, and verbs can require a change or mutation of their stem vocal in their inflected forms. Moreover, separable prefix verbs are preceded by a prefix, which has to be separated from the stem and put behind the verb in most cases. Therefore, a comprehensive German lexicon, enriching inflection rules with irregular forms, is needed for German surface realisation.

Unlike lexica for other tasks, a lexicon for NLG and specifically for surface realisation does not need to contain information about the semantics of a word, unlike e.g. GermaNet (Hamp and Feldweg, 1997). However, it is also not sufficient for such a lexicon to contain the different lemmata, it also has to include the inflections of each lemma. In German, a single noun lemma can have six additional inflections: genitive singular, genitive plural, dative singular, dative plural, accusative singular, and accusative plural form. For irregular verbs like "sein" (to be), this number can easily double.

So far, no well-structured and openly available lexicon of this kind exists for the German language. Creating such a resource from scratch would require a huge effort. However, with Wiktionary (cf. Section 3.), there exists a crowd-sourced resource which can form the base of such a lexicon, as we will show in this paper.

2. Other Lexica

SimpleNLG (Gatt and Reiter, 2009) is arguably the most popular rule-based open source surface realiser. Version 4.4.8 of SimpleNLG[1] comes with a default lexicon containing more than 6,000 lemmata and only very little information about inflection. Even irregular verbs like "be" only contain information about the past participle and simple participle.

The old German version of SimpleNLG (Bollmann, 2011) comes with what the author calls a "toy lexicon". This lexicon consists of only around 100 lemmata. The lexicon is based on the larger IMSLex from Lezius et al. (2000) which contains more than 50,000 lemmata of which about 11,000 are adjectives, 1,000 adverbs, 22,500 nouns, 300 particles, 10,000 proper nouns, and 6,000 verbs.[2] IMSLex contains information on inflection, word formation, and valence. Like MucLex, it also does not contain semantic information. However, the authors suggest that semantic information can be added from GermaNet.

Sennrich and Kunz (2014) build a German morphological lexicon for the morphological analyser SMOR (Schmid et al., 2004) from which is also based on information extracted from Wiktionary. The latest published version of their lexicon consists of 78,161 lemmata. For each lemma, Zmorge contains the stem, the part-of-speech, the origin and the SMOR inflection class. Tools like SMOR or Morphy (Lezius, 2000) could also be used to create or at least extend lexica, based on their morphological rules.

Currently, there exist six additional versions of SimpleNLG, for French, Italian, Spanish, Dutch, Mandarin, and Galician. Each of these versions comes with a lexicon for their respective language. Table 1 shows a comparison of the lexicon size of SimpleNLG in different languages.

There also exist multiple German lexica for the other popular open source surface realiser OpenCCG.[3] Vancoppenolle et al. (2011) provide a lexicon with approximately 250 lemmata. Hockenmaier (2006) used the TIGER corpus (Brants et al., 2004) to derive a German lexicon for OpenCCG containing more than 2,500 lemmata and more than 46,000 derived word forms.

With more than 100,000 lemmata (cf. Section 5.) and more than 670,000 word forms, MucLex is, to the best of our knowledge, the biggest open lexicon for the German language of its kind.

[1] https://github.com/simplenlg/simplenlg

[2] https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/imslex/

[3] https://github.com/OpenCCG/openccg

Language	Lemmata	Authors
English	6,000	Gatt and Reiter (2009)
German	100	Bollmann (2011)
French	3,000	Vaudry and Lapalme (2013)
Italian	35,000	Mazzei et al. (2016)
Spanish	76,000	Soto et al. (2017)
Dutch	79,000	de Jong and Theune (2018)
Mandarin	1,000	Chen et al. (2018)
Galician	11,000	Cascallar-Fuentes et al. (2018)

Table 1: Approximate number of lemmata in the standard lexicon of different SimpleNLG language versions

3. Wiktionary

Wiktionary[4] is a crowd-sourced open lexicon which is available in 130 languages. The German version contains more than 115,000 lemmata and has 176,431 registered users. Content from Wiktionary is dual-licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)[5] and GNU Free Documentation License (GFDL)[6]. We decided to use the CC BY-SA 3.0 license for our project, therefore MucLex is available under the same license.

The content of Wiktionary can be downloaded as XML dump, however, these dumps are only semi-structured. The lemma pages, which contain grammatical information like inflections, are structured with the wiki markup language. Listing 1 shows the relevant part of semi-structured entry for the verb "sein" (to be) from Wiktionary, Listing 2 shows the same information for the noun "Wortschatz" (wordstock) and Listing 3 for the adjective "schnell" (fast). Unfortunately, this format is not structured enough to be used as a lexicon for surface realisation directly. Therefore, we developed a parser which transforms the semi-structured Wiktionary XML dump into a well-structured XML-lexicon which can be easily digested by surface realisers like SimpleNLG.

```
{{Deutsch Verb Übersicht
|Präsens_ich=bin
|Präsens_du=bist
|Präsens_er, sie, es=ist
|Präteritum_ich=war
|Partizip II=gewesen
|Konjunktiv II_ich=wäre
|Imperativ Singular=sei
|Imperativ Plural=seid
|Hilfsverb=sein
}}
```

Listing 1: Semi-structured information for the verb "sein" in Wiktionary

```
{{Deutsch Substantiv Übersicht
|Genus=m
|Nominativ Singular=Wortschatz
|Nominativ Plural=Wortschätze
|Genitiv Singular=Wortschatzes
|Genitiv Plural=Wortschätze
|Dativ Singular=Wortschatz
|Dativ Singular*=Wortschatze
|Dativ Plural=Wortschätzen
|Akkusativ Singular=Wortschatz
|Akkusativ Plural=Wortschätze
}}
```

Listing 2: Semi-structured information for the noun "Wortschatz" in Wiktionary

```
{{Deutsch Adjektiv Übersicht
|Positiv=schnell
|Komparativ=schneller
|Superlativ=schnellsten
|Bild 1=Fast train (4712207733).jpg|
    ↪ thumb|1|ein ''schneller'' [[Zug]]
|Bild 2=VLI VL750 MStick Angle
    ↪ (4994860126).jpg|thumb|2|''
    ↪ schnelle'' [[üDatenbertragung]]
    ↪ mit einem [[USB-Stick]]
|Bild 3=Windflower-05237-nevit.JPG|
    ↪ thumb|3|sich ''schnell'' [[drehen
    ↪ ]]
}}
```

Listing 3: Semi-structured information for the adjective "schnell" in Wiktionary

Our parser traverses the XML dump from top to bottom. XML elements not representing word entries or inflection tables for verbs are skipped. Word entries include large amounts of information not relevant for rule-based NLG, such as word origin, synonyms, and pronunciation. In order to keep the lexicon small in file size, but as powerful as possible, we only keep the information that is necessary for the surface realisation.

The Wiktionary XML dump contains multiple XML elements for each verb. One of them is the main lemma, or base word entry, which is structured like the entries for other parts of speech, like nouns. This entry contains part of speech, some conjugation forms in present and preterite, the participle II form, and other information. However, it does not include information about whether the verb is regular, irregular or reflexive. This information can be found in verb inflection tables, which form a separate XML element outside the base word entry and are located in another position in the dump. The parser thus first checks whether an entry is an inflection entry or not, extracts the relevant information, examines if an entry for the same word already exists in the newly generated lexicon and if that is the case, adds new elements to this word, instead of creating a duplicate entry.

The lexicon currently excludes person names from Wiktionary, in order to not unnecessarily enlarge the lexicon

[4] https://www.wiktionary.org/

[5] https://creativecommons.org/licenses/by-sa/3.0/deed.en

[6] https://www.gnu.org/licenses/fdl-1.3.html

file. The parser is written in Python and published under the Mozilla Public License Version 2.0[7]. The code and the lexicon itself are available from https://github.com/sebischair/MucLex.

4. Format

The format of the lexicon is based on the format of the default English lexicon from SimpleNLG. The lexicon consists of a word entry for every lemma. Depending on their part of speech, these entries have different attributes. Common attributes for all words in the lexicon are the lemma (base), a unique identifier (id), and the part of speech (category).
Entries for nouns (cf. Listing 4) include the word's gender (genus) and its singular and plural forms for all grammatical cases (nominative, genitive, dative, accusative).

```
<word>
   <base>Wortschatz</base>
   <id>47</id>
   <category>noun</category>
   <plural>Wortschätze</plural>
   <genus>m</genus>
   <genitive_sin>
      Wortschatzes
   </genitive_sin>
   <genitive_pl>
      Wortschätze
   </genitive_pl>
   <dative_sin>
      Wortschatz
   </dative_sin>
   <dative_pl>
      Wortschätzen
   </dative_pl>
   <akkusative_sin>
      Wortschatz
   </akkusative_sin>
   <akkusative_pl>
      Wortschätze
   </akkusative_pl>
</word>
```

Listing 4: Data format for nouns

Verbs can appear in a vast amount of different conjugated forms in the German language. Incorporating all these forms for every tense, person, mood, and voice would unnecessarily enlarge the lexicon. Hence the lexicon only covers a set of conjugated forms broad enough to create all inflected forms in present, preterite, perfect, pluperfect, and future tenses, for indicative mood, and for active and passive voice by simple rules. This includes regular as well as irregular verbs. For verbs changing their stem in preterite tense, for instance, the preterite stem can be extracted from one included preterite form. All further forms can be built by rules using the given stem. Listing 5 shows the entry for the irregular verb "sein" (to be).

```
<word>
   <base>sein</base>
   <id>35</id>
   <category>verb</category>
   <regular>False</regular>
   <separable>False</separable>
   <reflexive>False</reflexive>
   <plFirstThirdPerPres>
      sind
   </plFirstThirdPerPres>
   <plSecPerPres>seid</plSecPerPres>
   <preterite>war</preterite>
   <participle2>gewesen</participle2>
   <firstPerPres>bin</firstPerPres>
   <secPerPres>bist</secPerPres>
   <thirdPerPres>ist</thirdPerPres>
</word>
```

Listing 5: Data format for verbs

Adjectives in the German language can occur in their base form, comparative, and superlative. Since the comparative and superlative can be build irregular, these forms are contained in the lexicon. An example entry is shown in Listing 6.

```
</word>
   <base>schnell</base>
   <id>3</id>
   <category>adjective</category>
   <comp>schneller</comp>
   <sup>schnellsten</sup>
</word>
```

Listing 6: Data format for adjectives

5. Size

In total, the MuxLex lexicon contains 101,509 distinct lemmata (cf. Table 2), which is slightly less than the total number contained in Wiktionary, because we excluded, for example, person names. With 78,780 lemmata, nouns are the biggest group, followed by adjectives with 11,456 lemmata, verbs with 10,289 lemmata, and adverbs with 1,127 lemmata. Overall, more than 670,000 word forms are included for these lemmata. Due to the relatively heavy XML structure, the file-size of the lexicon is 36 Megabytes. The XML dump of the German Wiktionary, in comparison, takes up more than 1.25 Gigabytes.

PoS	Lemmata
Nouns	78,780
Verbs	10,289
Adjectives	11,156
Adverbs	1,127
Total	101,509

Table 2: Amount of lemmata contained in MucLex by part of speech

6. Limitations

Although MucLex is much larger than existing lexica, it still has some limitations. For example, irregular subjunctive and imperative forms are currently not included in the

[7] https://www.mozilla.org/en-US/MPL/2.0/

lexicon but could be included in the future, because they are present in Wiktionary.

Moreover, only some common German compound nouns are currently included in the lexicon. Some nouns commonly used in compound words, for example "Test" ("test"), or "Haus" ("house"), offer large lists of compound words built from them in their Wiktionary entry, e.g. "Crashtest" ("crash test") or "Testfahrer" ("test driver"). But because these compound words do not have their own Wiktionary entry, they are currently not covered by the parser and therefore not part of the lexicon. Word entries with the base form and the separated parts used in a compound word could be added to the lexicon in order to increase the compound word coverage. However, it is debatable whether a lexicon should contain any compound nouns at all or whether it should just contain the base nouns and leave the composition to the surface realiser. There are existing approaches on how to automatically split compound words into their respective parts (e.g. by Baroni et al. (2002), Koehn and Knight (2003), Daiber et al. (2015), Sugisaki and Tuggener (2018), and Weller-Di Marco (2017)). The problem is far from being trivial and there are many compound nouns which behave "irregular" and would hence need to be included in the dictionary anyway. The most sensible solution might, therefore, be a mix of split approaches and a lexicon for irregular compound nouns.

If several inflected forms for a word are listed in Wiktionary, for example, "des Landes" and "des Lands" for the noun "Land" ("country"), the first form is taken. This might pose a limitation for nouns possessing several plural forms which have different meanings. The German word "Bank", for instance, may mean "bench" or "bank" (credit institute). In the plural form, the semantic difference becomes visible, as "benches" in German is "Bänke", but "banks" results in "Banken". The lexicon currently includes only the first-named plural form.

7. Conclusion

With more than 100,000 lemmata and 670,000 word forms, MucLex is currently the largest open German lexicon suitable for surface realisation. Since the German Wiktionary is constantly growing, MucLex will also grow in the future. By publishing the necessary parser to automatically create new versions of the lexicon, we want to ensure that all changes and additions to Wiktionary are available quickly for users of MucLex. Moreover, we hope to encourage people to contribute missing words to Wiktionary and hence not only contribute to our own lexicon but to openly available linguistic knowledge more broadly. In the future, we would like to address the limitations mentioned in Section 6. by improving the capabilities of the parser. MucLex is already used in conjunction with the new German version of SimpleNLG (Braun et al., 2019).

8. Acknowledgements

This work has been sponsored by the German Federal Ministry of Education and Research (BMBF) grant A-SUM 01IS17049 and Allianz SE.

9. Bibliographical References

Baroni, M., Matiasek, J., and Trost, H. (2002). Predicting the components of german nominal compounds. In *ECAI 2002: 15th European Conference on Artificial Intelligence*, pages 470–474.

Bollmann, M. (2011). Adapting simplenlg to german. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 133–138.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). Tiger: Linguistic interpretation of a german corpus. *Research on language and computation*, 2(4):597–620.

Braun, D., Klimt, K., Schneider, D., and Matthes, F. (2019). Simplenlg-de: Adapting simplenlg 4 to german. In *Proceedings of the 12th International Conference on Natural Language Generation*, Tokio, Japan. Association for Computational Linguistics.

Cascallar-Fuentes, A., Ramos-Soto, A., and Bugarın, A. (2018). Adapting simplenlg to galician language. *INLG 2018*, page 67.

Chen, G., van Deemter, K., and Lin, C. (2018). Simplenlg-zh: a linguistic realisation engine for mandarin. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 57–66.

Daiber, J., Quiroz, L., Wechsler, R., and Frank, S. (2015). Splitting compounds by semantic analogy. In *Proceedings of the 1st Deep Machine Translation Workshop*, pages 20–28, Praha, Czechia. ÚFAL MFF UK.

de Jong, R. and Theune, M. (2018). Going dutch: Creating simplenlg-nl. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 73–78.

Gatt, A. and Reiter, E. (2009). SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93, Athens, Greece, March. Association for Computational Linguistics.

Hamp, B. and Feldweg, H. (1997). Germanet-a lexical-semantic net for german. *Automatic information extraction and building of lexical semantic resources for NLP applications*.

Hockenmaier, J. (2006). Creating a CCGbank and a wide-coverage CCG lexicon for German. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 505–512, Sydney, Australia, July. Association for Computational Linguistics.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 187–193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lezius, W., Dipper, S., and Fitschen, A. (2000). Imslex representing morphological and syntactic information in a relational database. In *Proceedings of the 9th EU-*

RALEX International Congress, pages 133–139. Citeseer.

Lezius, W. (2000). Morphy-german morphology, part-of-speech tagging and applications. In *Proceedings of the 9th EURALEX International Congress*, pages 619–623. University of Stuttgart Stuttgart.

Mazzei, A., Battaglino, C., and Bosco, C. (2016). Simplenlg-it: adapting simplenlg to italian. In *Proceedings of the 9th International Natural Language Generation conference*, pages 184–192.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Sennrich, R. and Kunz, B. (2014). Zmorge: A german morphological lexicon extracted from wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14)*, pages 1063–1067, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Soto, A. R., Gallardo, J. J., and Diz, A. B. (2017). Adapting simplenlg to spanish. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 144–148.

Sugisaki, K. and Tuggener, D. (2018). German compound splitting using the compound productivity of morphemes. In *14th Conference on Natural Language Processing-KONVENS 2018*, pages 141–147. Austrian Academy of Sciences Press.

Vancoppenolle, J., Tabbert, E., Bouma, G., and Stede, M. (2011). A german grammar for generation in open ccg. In *Multilingual resources and multilingual applications: Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 145–150.

Vaudry, P.-L. and Lapalme, G. (2013). Adapting simplenlg for bilingual english-french realisation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 183–187.

Weller-Di Marco, M. (2017). Simple compound splitting for German. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 161–166, Valencia, Spain, April. Association for Computational Linguistics.