Augmenting a German Morphological Database by Data-Intense Methods

Petra Steiner
Friedrich-Schiller-Universität Jena
Jena, Germany
petra.steiner@uni-jena.de

Abstract

This paper deals with the automatic enhancement of a new German morphological database. While there are some databases for flat word segmentation, this is the first available resource which can be directly used for deep parsing of German words. We combine the entries of this morphological database with the morphological tools SMOR and Moremorph and a context-based evaluation method which builds on a large Wikipedia corpus. We describe the state of the art and the essential characteristics of the database and the context method. The approach is tested on an inflight magazine of Lufthansa. We derive over 5,000 new instances of complex words. The coverage for the lemma types reaches up to over 99 percent. The precision of new found complex splits and monomorphemes is between 0.93 and 0.99.

1 Introduction

German is a language with complex processes of word formation, of which the most common are compounding and derivation. Segmentation and analysis of the resulting word forms are challenging as spelling conventions do not permit spaces as indicators for boundaries of constituents as in (1).

(1) Verkehrsamt ‘tourist office’

For long orthographical word forms, many combinatorially possible analyses exist, though usually only one of them has a conventionalized meaning (see Figure 1). For instance, for *Verkehrsamt* ‘traffic office, tourist office’, word segmentation tools can yield the wrong split containing one with the smaller number of word tokens *Verkehr* ‘traffic’ and *Samt* ‘velvet’.

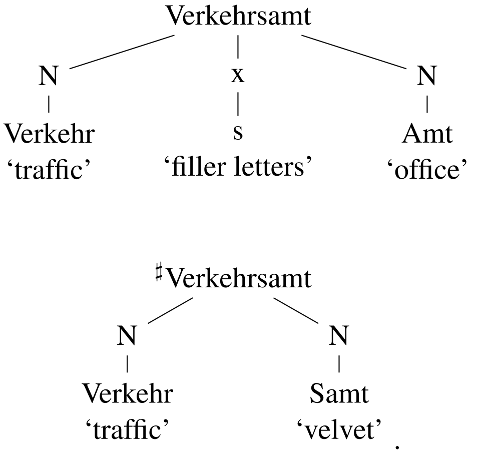

Figure 1: Ambiguous analysis of *Verkehrsamt* ‘tourist office’

In this case, there is a linking element within the word form which could be wrongly interpretated as part of a morph. Such elements function as morphophonological structure markers.[1]

German compounds can consist of derivatives, or compounds can be subject to further derivation. In (1), *Verkehr* is the result of a conversion process from *verkehren* ‘to run, to fly’, which again consists of a prefix and a verb stem (see Figure 2). On each level of morphological segmentation, the number of possible analyses is 2^n. This number can be reduced by excluding implausible constructions such as suffixes at the beginning of a construct. On the other hand, it has to be multiplied by the number of homonyms for the segmented forms. Therefore, automatic segmentations with more than ten possible analyses for one word are no rare case.

However, finding the correct segmentations and morphological structures is essential for terminologies and translation (memory) tools, information retrieval, and as input for tex-

[1] By some approaches, such linking elements are considered as a special kind of morphemes and called *Fugenmorpheme*. We like to avoid such classifications and use the labels *filler letters* or *interfix*.

Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 178–188
Florence, Italy. August 2, 2019 ©2019 Association for Computational Linguistics

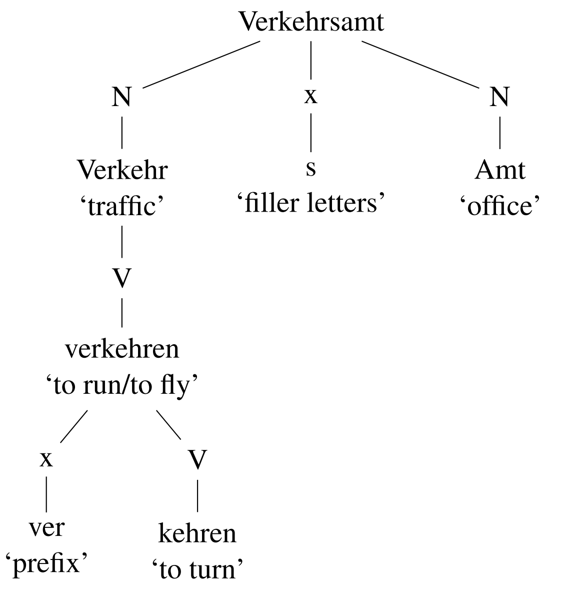

Figure 2: Complex analysis of *Verkehrsamt* 'tourist office'

tual analyses. Deep parsing of complex morphological structures produces disambiguation such as *(Fremde|n|verkehr)|s|amt* 'tourism office' instead of the tautological interpretation ♯*Fremde|n|(Verkehr|s|amt)* 'foreigner tourist office'.[2] Such analyses can help improving the quality of translation and retrieval tasks.

Moreover, counts of morphs, and morphological structures are useful for inducing hypotheses about statistical tendencies and quantitative laws, e.g. Menzerath's law (Cramer, 2005) or the Principle of Early Immediate Constituents (Hoffmann, 1999), which has not yet been corroborated for the word level by statistical tests.

In this paper, we will apply a hybrid approach for finding the correct splits of words and augmenting a morphological database. In Section 2, we provide a concise overview of previous work in word segmentation and word parsing for German. In Section 3, we introduce two linguistic tools we will be using later. *SMOR* is a well-known morphological tool. We describe how we modified its lexicon and exploited and changed its internal results by the add-on module *Moremorph*. Section 4 introduces our morphological database which was built on the basis of the linguistic databases CELEX and GermaNet. Section 5 describes the data-intense procedures for the morphological analyses and supervised database enhancements. In Section 6, we test our method on a corpus of an inflight journal. Finally, we discuss our results and give an outlook for future developments.

2 Related Work

The first developments in morphological segmentation tools for German date back to the Nineties. Most of them are based on finite state machines. Gertwol (Haapalainen and Majorin, 1995), MORPH (Hanrieder, 1996), Morphy (Lezius, 1996; Lezius et al., 1998) and later SMOR (Schmid et al., 2004) and TAGH (Geyken and Hanneforth, 2006) generate morphological analyses for complex German words, yielding results for derivatives and compounds. All these analyses are flat word splittings and often include dozens of segmentation versions.

There are different ways to tackle such kind of ambiguity, most of which are applied merely to compounds and yield flat segmentations of the immediate constituent level.

Cap (2014) and Koehn and Knight (2003) use ranking scores, such as the geometric mean, for the different morphological analyses and then choose the segmentation with the highest ranking.

Another approach consists in exploitation the sequence of letters, e.g. by pattern matching with tokens (Henrich and Hinrichs, 2011, 422) or lemmas (Weller-Di Marco, 2017). Ziering and van der Plas (2016) use normalization methods which are combined with ranking by the geometric mean. Ma et al. (2016) apply Conditional Random Fields modeling for letter sequences. Daiber et al. (2015) extract candidates of compound splits by string comparisons with corpus data.

Recent approaches exploit semantic information for the ranking of compound splittings. Riedl and Biemann (2016) utilize look-ups of similar terms inside a distributional thesaurus. Their ranking score is a modification of the geometric mean.

Ziering et al. (2016) use the cosine as a measure for semantic similarity between compounds and their hypothetical constituents and combine these similarity values by computing the geometric means and other scores for each produced split. The scores are then used as factors to be multiplied by the scores of former splits.

One of the few approaches tackling deep morphological analyses is Ziering et al. (2016). Their investigation considers left-branching compounds

[2] The complete structure of Fremdenverkehrsamt 'tourism traffic office, tourist office' is represented in Figure 4.

consisting of three lexemes. Their distributional semantic modelling often fails to find the correct binary split if the head is too ambiguous to correlate strongly with the first part. But in general, using the semantic context is a sensitive disambiguation method. Ziering and van der Plas (2016) develop a splitter which makes use of normalization methods and can be used recursively by reanalyzing the results of splits. Their evaluation is based on the binary compounds of GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2011).

Würzner and Hanneforth (2013) use a probabilistic context free grammar for full morphological parsing, but restrict their approach to derivational adjectives.

Most these approaches build upon corpus data. Only Henrich and Hinrichs (2011) enrich the output of morphological segmentation with information from the annotated compounds of GermaNet to disambiguate such structures. This can in a further step yield hierarchical structures but presupposes that the entries for the components exist inside the database. Steiner and Ruppenhofer (2018) build on this idea to derive more complex morphological structures from lexical resources. In 5, we come back to this and will exploit their resource.

3 SMOR: A Morphological Tool for German and its Add-On Moremorph

3.1 SMOR

SMOR is a widely used morphological segmentation tool (e.g. Cap (2014), Henrich and Hinrichs (2011), Steiner and Ruppenhofer (2015), Ziering et al. (2016)). It is based on two-level morphology (Koskenniemi, 1984) and implemented as a set of finite-state transducers. For German, a large set of lexicons is available. These lexicons contain information about inflection, parts of speech and classes of word formation, e.g. abbreviations and truncations. The tag set used is compatible with the STTS (Stuttgart Tübingen tag set, Schiller et al. (1995)).

SMOR produces different levels of granularity and different representation formats with different transducers and options. Example (2) and (3) show two simplified outputs of fine-grained analyses for *Verkehrsamt* 'traffic office, tourist office' and *Fremdenverkehrsamt* 'foreign-traffic office, tourist office'. For the sake of simplicity, we removed case and number.

(2) Verkehr<NN>Samt<+NN>
Verkehr<NN>Amt<+NN>
ver<VPREF>kehren<V>Samt<+NN>

(3) Fremdenverkehr<NN>Samt<+NN>
Fremdenverkehr<NN>Amt<+NN>
Fremd<Adj>verkehr<NN>Samt<+NN>
Fremd<Adj>verkehr<NN>Amt<+NN>
Fremd<Adj>ver<VPREF>kehren<V>
Samt<+NN>

In (2), the word form *Verkehrsamt* 'tourist office' is analyzed in three different ways, of which two show the erroneous interpretation of the string *samt* 'velvet' as a noun. (3) shows the same error in three of its five segmentations. The categories consist of parts of speech (<NN>, <V>) for free morphs and the position of bound morphemes (e.g. <VPREF> for 'verbal prefix').

3.2 Moremorph

While SMOR is a reliable foundation for the analysis of word forms which have not been found before, it comes with some small drawbacks. Moremorph aims at improving and adjusting the output of SMOR.

As can be seen from the second line of (2), the SMOR output does not indicate if there are filler letters (or interfixes) inside a word.

However, the information exists inherently in intermediate SMOR output which can be reanalyzed by Moremorph. Therefore, filler letters (FL) can be marked as in (4):

(4) Verkehr s Amt NN FL NN <NN>

This annotation shows the morphs on the lexical level, their classes with filler letters, and finally the part of speech of the word form in angle brackets. (5) presents the Moremorph representation of (3). In the last three analyses, there is one tag more than the number of splits due to the noun conversion of *fremd* 'foreign' to *Fremde* 'foreigner'.

(5) a. Fremdenverkehrsamt
Fremdenverkehr Samt
NN NN <NN>

b. Fremdenverkehrsamt
Fremdenverkehr s Amt
NN FL NN <NN>

c. Fremdenverkehrsamt
fremd en Verkehr Samt
ADJ NNSUFF FL NN NN <NN>

d. Fremdenverkehrsamt
fremd en Verkehr s amt
ADJ NNSUFF FL NN FL NN <NN>

e. Fremdenverkehrsamt
fremd en ver kehr Samt
ADJ NNSUFF FL VPREF V NN <NN>

Moremorphs uses SMOR lexicons which we adapted to the current task. The original version of the names lexicon comprised 14,998 entries, the final extended version 16,718 entries. During the project, the lexicon was constantly extended and cleaned and its entries were revised. The final version used for the current work comprises 42,205 entries. Many changes of the rule sets were made in cooperation with Helmut Schmid according to our suggestions. For example, we changed the sets of characters or added adverbs as possible tag class for numbers. Other changes include the derivation of adjectives from names of location. Some of the finite-state transducers had to be changed for this.

We also standardized inconsistent analyses for orthographical variants with and without hyphenations and added some more special characters to the inventory of word structuring means.

This leads to consistent analyses for orthographical variants such as in (6). Also word forms with some other special characters not covered by SMOR can be processed now, as in (7).

(6) a. Flughafen Köln-Bonn 'Airport Cologne-Bonn'
b. Flughafen Köln/Bonn 'Airport Cologne/Bonn'.

(7) "Team Lufthansa"-Partner

(8) shows the output for (6-b) with the structuring character tagged as HYPHEN.

(8) Köln/Bonn Köln / Bonn
NPROP HYPHEN NPROP <NPROP>

4 A Lexical Database with Deep-Level Morphological Information

While most morphological analyzers build on the results of word splitters, we decided to take up a hybrid approach which combines the reliable entries of a morphological database with the augmented and further processed analyses of SMOR and Moremorph. Here, also another morphological tool could be chosen.

The German morphological tree database extracts its entries from a. the refurbished CELEX database (Baayen et al., 1995; Steiner, 2016) for German morphology (Burnage, 1995; Gulikers et al., 1995) and b. the compound analyses from the GermaNet database (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2011; Steiner, 2017). For both preprocessed datasets, the derivation of complex structures was performed recursively, by combining the GermaNet analyses with the analyses from CELEX.

The tree building tool provides different parameters for the analysis. We chose to enrich the data with information on diachronic derivation and permitted a depth of six levels for the morphological analyses. (9) shows the morphological structures for (9-a) *Verkehrsamt* 'tourist office', (9-b) *Verkehrsanlage* 'traffic facility', and (9-c) *Verkehrsbehinderung* 'traffic obstruction'. (9-b) comprises diachronic derivational information, showing the noun *Anlage* 'facility/lay out' as derived from the verb *anlegen* 'lay out'.

```
(9)  a.  Verkehrsamt
          (*Verkehr*
            (*verkehren*
              ver|
             kehren))|
          s|
          Amt

     b.  Verkehrsanlage
           (*Verkehr*
            (*verkehren*
              ver|
              kehren))|
          s|
          (*Anlage*
            (*anlegen*
              an|
              legen))
```

c. Verkehrsbehinderung
(*Verkehr*
(*verkehren*
ver|
kehren))|
s|
(*Behinderung*
(*behindern*
be|
hindern)|
ung)

The number of entries for this databases of the morphological trees amounts to 101,588. In addition, we extracted 6,339 types of monomorphemes from the refurbished German CELEX database.

5 Combining Morphological Databases with a Segmenter

In the following, we combine the morphological database with a morphological segmenter and a contextual evaluation process. If the database look-up fails, the time-consuming word splitting and evaluation is started. Then the output of Moremorph is analyzed by a contextual method by exploiting a very large corpus. If this fails, frequencies counts of a very large corpus is the back-off strategy. The new analyses are added to a set of new splits.

At the end of each word analysis, all subparts of the word are being searched within the database and the newsplit set. This leads to incrementally more fine-grained entries.

Figure 3 presents an overview. It shows two databases of morphological trees: the German morphological tree database and a incremental database for all newly found morphological analyses. Furthermore, it comprises a set of monomorphemes.

5.1 Basic Look-Up

As shown in Figure 3, a look-up finds the respective tree or the simplex form for the word within the lexicons. Before this is added to the results, all of its subparts are being looked up within the databases and the new splits. These subanalyses are being integrated to its new analysis. Old entries within the lexical databases are being substituted for the new ones.

5.2 Finding Splits

If neither an entry inside the tree lexicons nor in the list of monomorphemes can be found, the Moremorph analyses are taken as the start for the further analysis. For each analysis, e.g. the five different ones of example (5), every possible combination of subtrees has to build. Some of them can be filtered out, because they are linguistically implausible, e.g. when a hypothetical subpart finishes with a prefix.

All plausible combinations of strings and tags undergo a contextual analysis, if occurrences for all subparts can be found within at least one text of the large corpus. Otherwise, a procedure of using the overall document frequencies together with a back-off strategy will be invoked.

5.2.1 Morphological Segmentation based on Contextual Information

For (unknown) compounds, we presuppose that each component can be found within the same close environments. Therefore, the frequencies of components in texts should be much lower for erroneous splits than the frequencies for correct segmentations.

We chose a large set of texts for the retrieval: the freely available and annotated German Wikipedia Korpus of 2015 (Margaretha and Lüngen, 2014).[3] We restricted ourselves to the 1.8 million texts subcorpus of the articles. The corpus was tokenized by a modified version of the tool from Dipper (2016) and lemmatized by the TreeTagger (Schmid, 1999). Text indices were built both for tokenized and lemmatized forms. For each text, all frequencies of lemmas and tokens were stored.

For each morphological split of a word form *wf* (sp$_{wf,n}$), the intersection of all texts comprising the word form *wf* and their hypothetical components $c_{wf,sp,1..n}$ is retrieved from the text indices. For every text t which includes all components for the word form wf ($c_{wf,sp,1}...c_{wf,sp,n}$) of a morphological split, the document frequencies (*df*) of the components are being retrieved and added to the sum of text frequencies score (*Stf*). For every hypothetical analysis, the highest value is chosen and the morphological analysis with this score is stored (Equation 1).

[3] see http://www1.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html

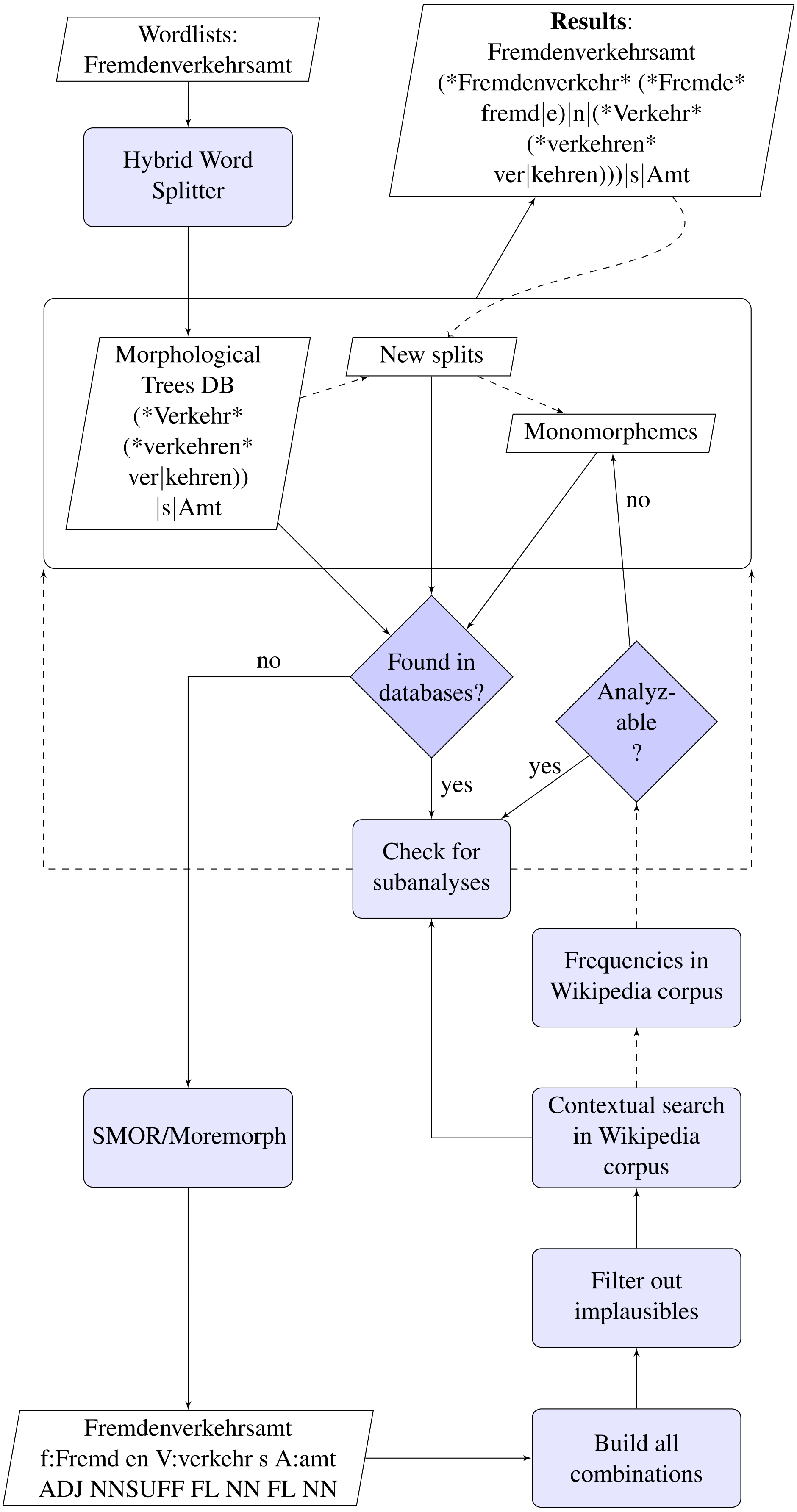

Figure 3: Hybrid word analysis: Morphological trees database, word segmenter, and two different evaluation procedures as alternative methods for word splitting

$$Best - Stf_{wf,sp,t} = \max_{1,t} \sum_{c_{wf,sp,1}}^{c_{wf,sp,n}} df_{1,n} \qquad (1)$$

Finally, for every hypothetical analysis, the highest value is chosen and the morphological analysis with this score is processed and stored.

5.2.2 Morphological Segmentation based on Document Frequencies

In case that no text can be found which includes the word form *wf* and the components of any of the hypothetical analyses, the corpus itself is considered as a textual enviroment in the widest sense. For each split, the sum of frequencies are being calculated. The hypothetical analysis with the highest value is chosen and the morphological analysis with this score is processed for the storage. In all other cases, the analysis will yield the hypothetical analysis of a monomorpheme.

5.3 Substitution of Analyses

Whenever an analysis by the $Best - Stf$ score or another look-up has been found, the analyses for its immediate constituents are being searched in the databases. By this, the lexicons can be incrementally enlarged and enriched. Figure 4 shows an example from our test corpus, which we used for the evaluation in Section 6.

The results are added to a database of new splits and can be added to the previous database after an evaluation.

6 Evaluation

6.1 Data

For testing the performance, we use *Korpus Magazin Lufthansa Bordbuch (MLD)* which is part of the DeReKo-2016-I (Institut für Deutsche Sprache, 2016) corpus[4]. It is an in-flight magazine with articles on traveling, consumption and aviation. For the tokenization, we enlarged and costumized the tokenizer by Dipper (2016) for our purposes. Multi-word units were automatically identified based on the multi-word dataset which we had augmented before. The resulting data comprises 276 texts with 5,202 paragraphs, 16,046 sentences and 260,114 tokens. The number of word-form types is 38,337. We are analyzing the lemmatized version of this corpus which was produced by the TreeTagger (Schmid, 1999), it comprises 27,902 lemma types.

6.2 Results

6.2.1 Coverage

15,622 lemma types can be found within the database. 12,280 lemma types are not covered by the databases, so they were re-analyzed by SMOR/Moremorph. We manually checked the results for the first 1,000 lemma types which could not be found in the database. Very often, these are derivatives, rare or nounce words, proper names or words containing proper names as in (10).

(10) a. ordnend 'ordering, regulatory'
b. Paris-Erfahrung 'Paris experience'
c. Winterspaß 'winter fun'

The details of the check against the German tree database are included in Table 1, with a coverage of 55.99% for the lemma types. This direct lookup saves a lot of computational effort. According to the quality of the database which is based on GermaNet and CELEX, the recall is extremely close to these numbers.

The remaining 44.01% of all lemma types were evaluated in the following way: We checked every split of the first thousand analyzed words. For ambiguous analyses, we accepted those which included a monomorphemic and a correct derivational analysis, as in (11), with (11-a) showing the segmentation of verb stem and derivational suffix.

(11) a. ordnend ordn end V PPres ADJ-SUFF <ADJ>
b. ordnend ordnend V <V>

If one or more splits were erroneous, as in (12-a), the analysis was rejected.

(12) a. Winterspaß Winter spaß NN NN <NN> 'winter fun'
b. ♯Winterspaß Winter s paß NN FL NN <NN> 'winter|s, filler letter| pass/passport'

We found 26 wrongly segmented words inside the sample of a thousand words from the SMOR/Moremorph output. This shows a good quality of the analysis. However, unknown

[4] See Kupietz et al. (2010) and http://www1.ids-mannheim.de/kl/projekte/korpora/archiv/mld.html for further information.

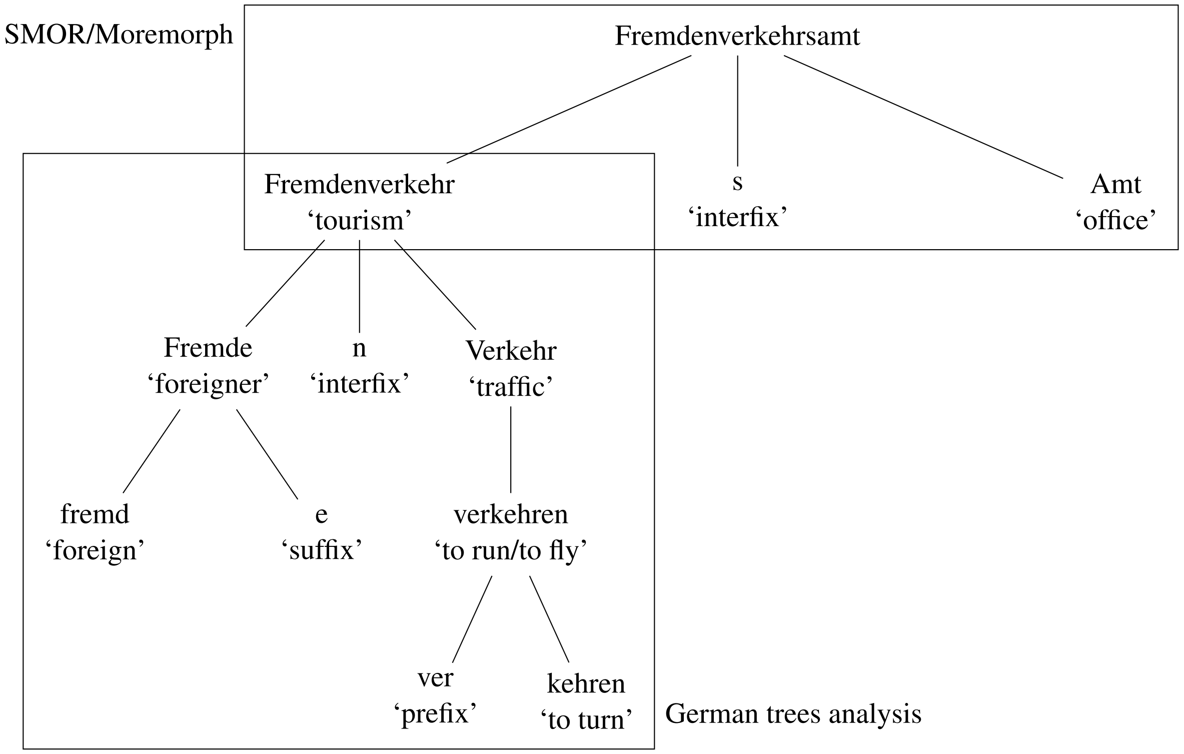

Figure 4: Database look-up and SMOR/Moremorph: Morphological analysis of *Fremdenverkehrsamt* ‘tourist office’

types were re-analysed as hypothetical monomorphemes during the further analysis. Often, these were names of airplane types or similar expressions. Therefore, the number of analyzed lemma types (27,902) corresponds to a full coverage. SMOR/Moremorph on its own was able to process 13,461 lemmas, the rest was classified as unknown. This good coverage is a direct result of the adjustment of the lexicons, which we described in 3.2, especially concerning the names lexicons.

	lemma types	**corpus size**
MLD corpus	27,902	260,114

	lemma types	**coverage**
Tree DB + monomorphs	15,622	55.99%
+ SMOR & Moremorphs	27,902	100%

Table 1: Coverage of DBs and SMOR analyses

6.2.2 Precision

The complete analyses of the hybrid morphological parsing yield 5,307 entries in the newsplit database and 5,973 new entries inside the monomorphemes. We analyzed the first 1,000 entries of the newly found splits and the first 2,000 entries within the monomorpheme set. Of the first set, we found 65 wrongly or imperfectly analyzed word forms. Most of them are three-part compounds such as (13) whose correct components were not found within a text. The morphemes were identified, but the ambiguity could not be resolved.

(13) (Berg|Regen|Wald) ‘mountain rain forest’

Another error are wrong analyses of derivative nouns which starts with a verb particle such as (14-a) , which is a derivative form of *anfahren* ‘to approach’ (14-b) and not a compound of *an* ‘at, to’ and *Fahrt* ‘ride’. There is a systematic mistake here which is caused by the high frequency of the first part which is usually a homograph of a preposition.

(14) a. ♯An|(*Fahrt* fahren|t) ‘approach’
b. (*anfahren* an|fahren)|t ‘approach’

The set of monomorphs comprise many new complex numbers and proper names. All of them were correctly included. Only three assignments are questionable. However, as these are proper names such as *Anneliese* which consists of two proper names *Anna* and *Liese*, and/or the analysis in CELEX was monomorphemous too (as for *Allerheiligen* 'All Saints'), the quality is very high. Therefore, the precision can be considered as high for this test corpus: 0.935 for new splits and 0.998 for newly found monomorphs.

6.3 Discussion

The results for the first hybrid deep-level morphology analyzer are promising. However, the errors concerning verb particles are systematic. They can be explained by the high frequency of verb particles in texts, which are often homographs of a preposition. For future research, we plan an adjustment by a factor which takes into account the relationship between word length in characters and word frequency as observed by Zipf and others (Prün, 2005). Köhler (1986) derives this relationship by a synergetic model. He corroborates the functional connection between the frequency classes of words and their average length. A measure directly derived from this function would penalize word segmentations with small morphemes and assign more weight to longer (and rare) components.

7 Conclusion and Outlook

This paper demonstrates how updating and exploiting linguistic databases for morphological analyses can be performed. By simple look-up, we reached a coverage of 56% of lemma types. As both underlying databases, CELEX and GermaNet, were manually revised, we can speak of very reliable analyses. The remaining unanalyzed words can be mostly covered by a conventional word segmenter after adjusting its lexicons. These analyses have a flat structure and undergo a procedure of constructing all combinations of possible analyses and a context-based search for the hypothetical constituents in a large corpus. The results for the lemma types are very promising: Over 99% of all words were covered by the combined morphological analyses.

New morphological analyses from the tree-building process can be added to the German tree database after a process of careful evaluation and selection.

The direction of the future research is therefore straightforward: it will lead towards creating complex analyses out of existing ones and augmenting the lexical databases.

8 Acknowledgement

Work for this publication was partially supported by the German Research Foundation (DFG) under grant RU 1873/2-1. I would like to thank the reviewers for their valuable feedback and my colleague Josef Ruppenhofer for making this work possible.

References

Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1995. The CELEX lexical database (CD-ROM).

Gavin Burnage. 1995. CELEX: A Guide for Users. In Harald Baayen, Richard Piepenbrock, and Léon Gulikers, editors, *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, Philadelphia, PA.

Fabienne Cap. 2014. *Morphological processing of compounds for statistical machine translation*. Ph.D. thesis, Universität Stuttgart.

Irene M. Cramer. 2005. Das Menzerathsche Gesetz (Menzerath's law). In Reinhard Köhler, Gabriel Altmann, and Raĭmond Genrikhovich Piotrovskiĭ, editors, *Quantitative Linguistik / Quantitative Linguistics - Ein internationales Handbuch / An International Handbook*, pages 659–687. M. de Gruyter.

Joachim Daiber, Lautaro Quiroz, Roger Wechsler, and Stella Frank. 2015. Splitting compounds by semantic analogy. In *Proceedings of the 1st Deep Machine Translation Workshop*, pages 20–28. ÚFAL MFF UK.

Stefanie Dipper. 2016. Tokenizer for German.

Alexander Geyken and Thomas Hanneforth. 2006. TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In *Finite State Methods and Natural Language Processing. 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002, pages 55–66. Springer.

Léon Gulikers, Gilbert Rattink, and Richard Piepenbrock. 1995. German Linguistic Guide. In Harald Baayen, Richard Piepenbrock, and Léon Gulikers, editors, *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, Philadelphia, PA.

Mariikka Haapalainen and Ari Majorin. 1995. GERTWOL und morphologische Disambiguierung für das

Deutsche. In *Proceedings of the 10th Nordic Conference on Computational Linguistics, Helsinki, Finland*.

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

Gerhard Hanrieder. 1996. MORPH - Ein modulares und robustes Morphologieprogramm für das Deutsche in Common Lisp. In Roland Hauser, editor, *Linguistische Verifikation Dokumentation zur Ersten Morpholymics 1994*, pages 53–66. Niemeyer, Tübingen.

Verena Henrich and Erhard Hinrichs. 2011. Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 420–426. Association for Computational Linguistics.

Christiane Hoffmann. 1999. Word order and the Principle of "Early Immediate Constituents" (EIC). *Journal of Quantitative Linguistics*, 6(2):108–116.

Institut für Deutsche Sprache. 2016. Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2016-I (Release from 31.03.2016).

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics-Volume 1*, pages 187–193. Association for Computational Linguistics.

R. Köhler. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*, volume 31 of *Quantitative linguistics*. Brockmeyer.

Kimmo Koskenniemi. 1984. A general computational model for word-form recognition and production. In *Proceedings of the 10th International Conference on Computational linguistics*, pages 178–181. Association for Computational Linguistics.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German reference corpus DeReKo: A primordial sample for linguistic research. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1848–1854, Valletta, Malta. European Language Resources Association (ELRA).

Wolfgang Lezius. 1996. Morphologiesystem Morphy. In *Linguistische Verifikation. Dokumentation zur ersten Morpholympics 1994*, pages 25–35. Niemeyer.

Wolfgang Lezius, Reinhard Rapp, and Manfred Wettler. 1998. A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for German. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference.*, pages 743–748.

Jianqiang Ma, Verena Henrich, and Erhard Hinrichs. 2016. Letter Sequence Labeling for Compound Splitting. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 76–81, Berlin, Germany. Association for Computational Linguistics.

Eliza Margaretha and Harald Lüngen. 2014. Building linguistic corpora from wikipedia articles and discussions. *JLCL*, 29(2):59–82.

Claudia Prün. 2005. Das Werk von G. K. Zipf (The work of G. K. Zipf). In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski, editors, *Quantitative Linguistik / Quantitative Linguistics - Ein internationales Handbuch / An International Handbook*, pages 142–152. DeGruyter.

Martin Riedl and Chris Biemann. 2016. Unsupervised Compound Splitting With Distributional Semantics Rivals Supervised Methods. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 617–622. Association for Computational Linguistics.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1995. Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, and Seminar für Sprachwissenschaft, Universität Tübingen.

Helmut Schmid. 1999. Improvements in Part-of-Speech Tagging with an Application to German. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 13–25. Springer Netherlands, Dordrecht.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 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)*. European Language Resources Association (ELRA).

Petra Steiner. 2016. Refurbishing a Morphological Database for German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Petra Steiner. 2017. Merging the trees. building a morphological treebank for German from two resources. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, January 23–24, 2018*, TLT'16, pages 146–160, Prague, Czech Republic.

Petra Steiner and Josef Ruppenhofer. 2015. Growing trees from morphs: Towards data-driven morphological parsing. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2015, University of Duisburg-Essen, Germany, 30th September - 2nd October 2015*, pages 49–57.

Petra Steiner and Josef Ruppenhofer. 2018. Building a Morphological Treebank for German from a Linguistic Database. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3882 – 3889, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Kay-Michael Würzner and Thomas Hanneforth. 2013. Parsing morphologically complex words. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing, FSMNLP 2013, St. Andrews, Scotland, UK, July 15-17, 2013*, pages 39–43.

Patrick Ziering, Stefan Müller, and Lonneke van der Plas. 2016. Top a splitter: Using distributional semantics for improving compound splitting. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 50–55, Berlin, Germany. Association for Computational Linguistics.

Patrick Ziering and Lonneke van der Plas. 2016. Towards Unsupervised and Language-independent Compound Splitting using Inflectional Morphological Transformations. In *Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 644–653. Association for Computational Linguistics.