Combining Data-Intense and Compute-Intense Methods for Fine-Grained Morphological Analyses

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

Abstract

This article describes a hybrid approach for German derivational and compositional morphology. Its first module is based on retrieval from morphological databases. The second module builds on the results of a word segmenter and uses a context-based approach by exploiting 1.8 million texts from Wikipedia for the disambiguation of multiple morphological splits. Insights from Quantitative Linguistics help countering two sparse-data problems. The results can become more fine-grained during each cycle of computation and be added to the lexical input data with or without supervision. The evaluation on an inflight magazine shows a good coverage and an accuracy of 93% for the deep-level analyses.

1 Introduction

German is a language with highly productive and complex processes of word formation. Moreover, spelling conventions do not permit spaces as indicators for boundaries of constituents as in (1). Therefore, the automatic segmentation and analysis of the resulting word forms are challenging.

(1) Arbeitsaufwand ‘work effort, expenditure of labor’

Often, many combinatorially possible analyses exist, though usually only one of them has a conventionalized meaning (see Figure 1). For instance, for *Aufwand* ‘expense, expenditure’, word segmentation tools can yield the wrong split. In this case, there is a linking element within the word form which could be wrongly interpretated as part of a morph.[1] Here, the wrong alignment leads to a result containing *Sauf* ‘to drink$_{animal}$, to booze’ and *Wand* ‘wall’ as erroneously segmented morphs.

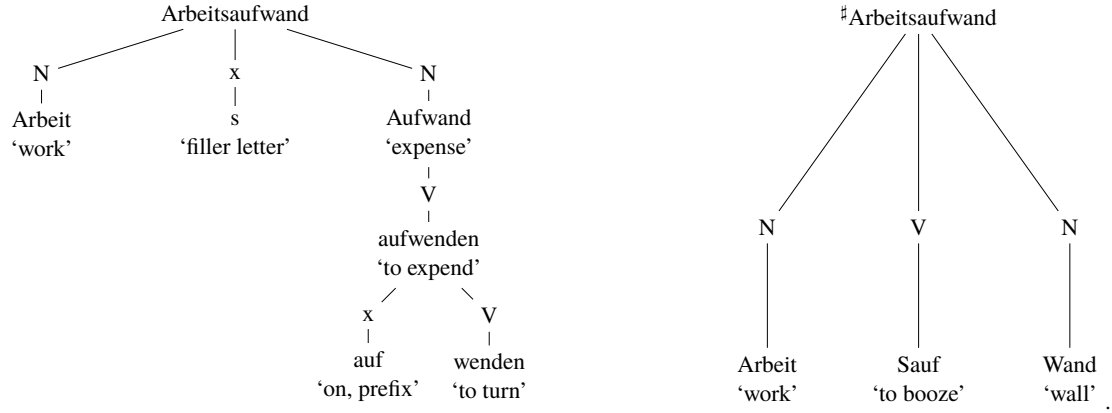

Figure 1: Ambiguous analysis of *Arbeitsaufwand* ‘expenditure of labor’

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

Proceedings of the 2nd Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019), pages 45–54,
Prague, Czechia, 19-20 September 2019.

German compounds can consist of derivatives and derivatives can have compounds as their constituents. In (1), *Aufwand* is the result of a conversion process from *aufwenden* 'to expend, to spend', which again consists of a prefix and a verb stem. Many orthographical words are highly ambiguous. Therefore, automatic segmentation with many possible analyses for one orthographical form is a standard problem for German word formation.

In this paper, we use a hybrid approach for finding the correct splits of words and augmenting a morphological database. In Section 2, we provide an overview of related work in word segmentation and word parsing for German with a focus on structural analysis. Section 3 describes the combination of data-intense procedures for the morphological analyses and supervised database enhancements with compute-intense methods by exploiting a Wikipedia corpus. We also derive some quantitative heuristics to cope with sparse data problems. Section 4 shows an evaluation of the analyses of ambiguous word forms from the corpus of an inflight journal. Section 5 summarizes the main points of our results, and finally Section 6 gives a short outlook on future work.

2 Related Work

The first developments in morphological segmentation tools for German started in the early Nineties. Most of them are based on finite state transducers, for instance Gertwol (Haapalainen and Majorin, 1995), Morphy (Lezius, 1996), and later SMOR (Schmid et al., 2004) and TAGH (Geyken and Hanneforth, 2006). These morphological segmenters for complex German words often include dozens of flat word splittings for derivatives and compounds. There are different ways to resolve 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. Sugisaki and Tuggener (2018) use a probabilistic model for composition activity of the constituents of noun compounds. They exploit the frequencies of large corpora.

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.

Recent approaches exploit semantic information for the ranking of compound splittings, such as look-ups of similar terms inside a distributional thesaurus such as Riedl and Biemann (2016). 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 weights to be multiplied by the scores of former splits. Their investigation considers left-branching compounds consisting of three lexemes by using distributional semantic modelling. If the head is too ambiguous to correlate strongly with the first part, this often fails. Here, the test data of the left-branching compounds is preselected.

Few approaches take steps into the direction of hierarchical word segmentation: Ziering and van der Plas (2016) develop a splitter which makes use of normalization methods and can be used recursively by re-analyzing the results of splits. Schmid (2005) tests the disambiguation of morphological structures by a head-lexicalized probabilistic context-free grammar. The input are flat segmentations from SMOR. The baseline of 45.3% for the accuracy is obtained by randomly selecting an analysis from the least complex results. The parser is trained by using the Inside-Outside algorithm on different models e.g. frequencies of tokens vs. types, and lexicalized vs. unlexicalized training and combinations of these during the iteration process. The best results reach 68%. Besides this, the paper systematically describes the pitfalls of automatic morphological segmentation for immediate constituents and morphs. Some of these cases will be reconsidered when discussing the test data in Section 4. Another advance with a probabilistic context free grammar for full morphological parsing was undertaken by Würzner and Hanneforth (2013), however, it is restricted 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. This can in a further step yield hierarchical structures. Steiner and Ruppenhofer (2018) and Steiner (2017) build on this idea to derive complex morphological structures from lexical resources. In the following section, we will describe how we use the combined morphological information of GermaNet and CELEX as the foundation for a hybrid analyzing tool.

3 Combining Contextual Retrieval with Data-intense Methods

We will combine the look-up in a morphological database with a morphological segmenter and a contextual evaluation process. Figure 2 presents an overview of the procedure. It shows two databases of morphological trees: the German morphological trees database and a incremental database for all newly found morphological analyses (new splits). Furthermore, it comprises a set of monomorphemic lexemes.

If the database retrieval fails, the word splitting and weighting of alternative morphological structures is started. The output of a segmentation tool is analyzed by a contextual method by exploiting 1.8 million texts. If this fails, frequencies counts of a very large corpus is the next strategy. Some effects of typical frequency distributions are compensated by adequate weights. At the end of each word analysis, all subparts of the word are being searched within the database and the newsplit set. In case that entries for these subparts exist, their analyses can integrated into the results. The deeper analysis can then substitute the former one or is added as a new entry. These possible changes are represented by the dashed lines between the check block and the lexical databases.

3.1 Data-intense Methods for Morphological Analysis

We use the German morphological trees database built by the tools of Steiner (2017). It combines the analyses of the German part of the CELEX database (Baayen et al., 1995) and the compound analyses from the GermaNet database (Henrich and Hinrichs, 2011). (2) shows the entry for (1) *Arbeitsaufwand*. It comprises information on compounding from GermaNet and derivation from CELEX, for instance the constituent *Aufwand* 'expenditure' is analyzed as a derivation from the verb *aufwenden* 'to expend'. Non-terminal constituents are marked by asterisks.

(2) Arbeitsaufwand (*Arbeit* arbeiten)|s|(*Aufwand* (*aufwenden* auf|wenden))

Figure 2 shows two databases of morphological trees: the German morphological trees database comprising 101,588 entries of complex lexemes, and an incremental database for all newly found morphological analyses. Furthermore, it comprises a set of monomorphemic lexemes, starting with 6,339 entries from the refurbished German CELEX database.

The hybrid word analyzer starts with a basic look-up. If this search can retrieve the respective tree or simplex form for the word, all of its subparts are being looked up within the lexical databases. These subanalyses are being integrated and old entries within the lexical databases are being substituted for the new ones. No further analyses are necessary.

3.2 Word Splitting and Contextual Retrieval

If neither an entry within the tree lexicons nor within the list of monomorphemes can be found, we use the output from SMOR as start for the further processing. We adjusted the SMOR output to our needs by the add-on Moremorph as it is described in Steiner and Rapp (in press). The flat structures include filler letters and tags of free and bound morphemes as in (3) for *Chefredakteurin* 'editor-in-chief$_{female}$'. The *l:s* notation shows lexical and surface characters of the two-level morphology which SMOR is based on. Eight of the ten analyses comprise false analyses, such as ♯(Chef|reden|Akte|Urin) '(Chief|to talk|file|urine)', all are analyzed as nouns (<NN>). We previously adjusted and enriched SMOR's lexicons and rules to our needs (Steiner and Rapp, in press). Under these premises and if inflectional information is cut off, the flat segmentation yields approximately two structures per word form. If case and number features were included, the number of analyses was about thrice as large (Schmid, 2005).

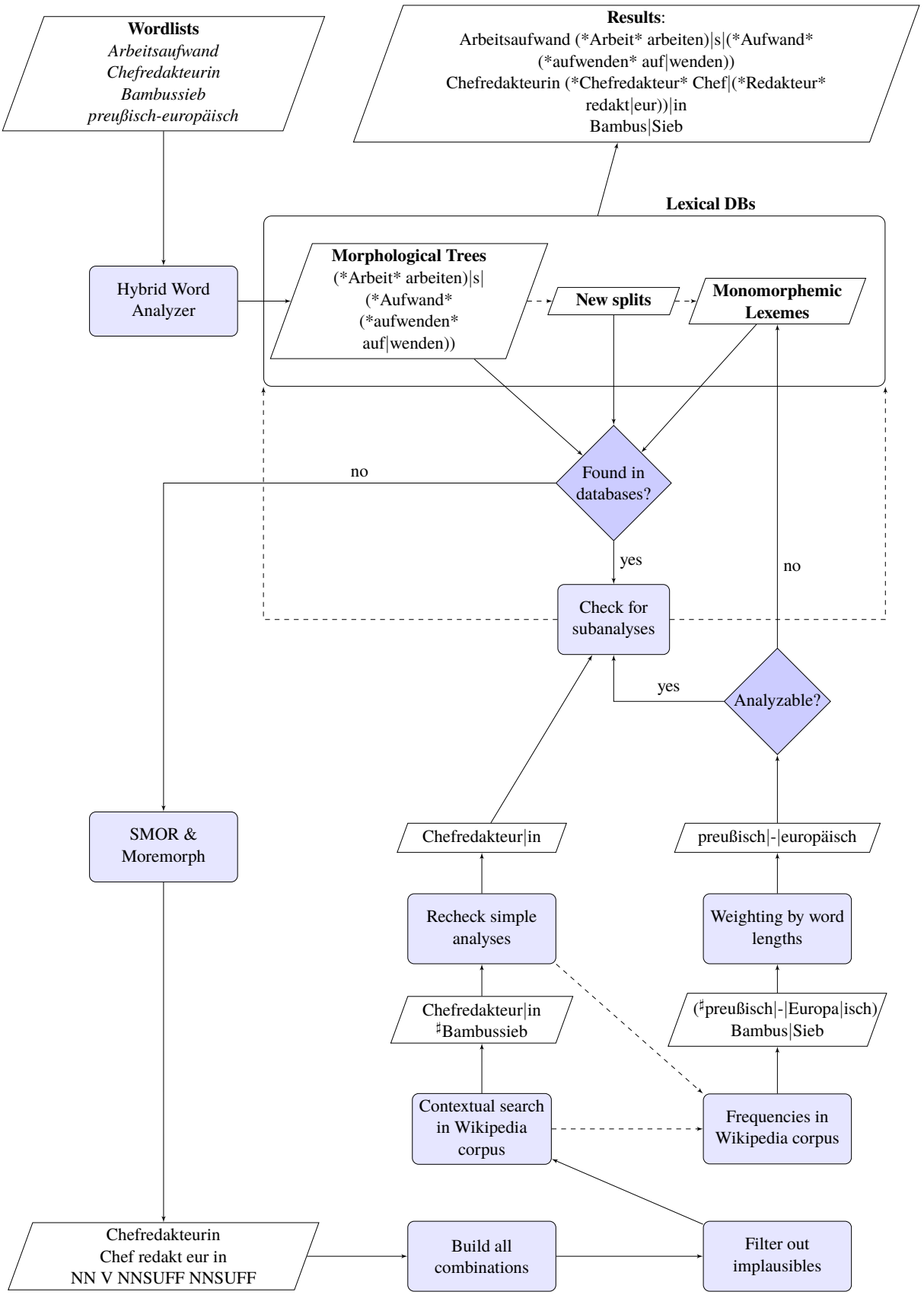

Figure 2: Hybrid word analysis: Morphological trees database, word segmenter, and two different evaluation procedures as alternative methods for word splitting

For each of these flat analyses, 2^{n-1} combinations of the immediate constituents exist, some of which are implausible, e.g. a constituent consisting just of a suffix. For the first flat analysis in (3), three splits are combinatorially possible and linguistically plausible (4). (4-a) describes a derivation of *Chefredakteur* 'editor-in-chief$_{masc}$' and the suffix *in*. But also, a composition of *Chef* and *Redakteurin* 'editor$_{female}$' is possible as in (4-b). Finally, (4-c) describes the erroneous analysis as a monomorphemic lexeme. Only the first combination shows the correct morphological structure.

(3)

Chefredakteurin	Chef R:redakteur in	NN NN NNSUFF	<NN>
Chefredakteurin	Chef rede:<>n:<> A:akte U:urin	NN V NNSUFF NN NN	<NN>
Chefredakteurin	Chef rede:<>n:<> A:akteur in	NN V NNSUFF NN NNSUFF	<NN>
Chefredakteurin	Chef rede:<>n:<> A:akt e U:urin	NN V NNSUFF NN FL NN	<NN>
Chefredakteurin	Chef rede:<>n:<> akt eur in	NN V NNSUFF V NNSUFF NNSUFF	<NN>
Chefredakteurin	Chef rede:<>n:<> A:akte U:urin	NN V NN NN	<NN>
Chefredakteurin	Chef rede:<>n:<> A:akteur in	NN V NN NNSUFF	<NN>
Chefredakteurin	Chef rede:<>n:<> A:akt e U:urin	NN V NN FL NN	<NN>
Chefredakteurin	Chef rede:<>n:<> akt eur in	NN V V NNSUFF NNSUFF	<NN>
Chefredakteurin	Chef redakt eur in	NN V NNSUFF NNSUFF	<NN>

(4)
a. [[NN,NN],[NNSUFF]] Chefredakteur|in
b. ♯[[NN],[NN, NNSUFF]] Chef|redakteurin
c. ♯[[NN, NN, NNSUFF]] Chefredakteurin

For all such flat analyses, all plausible combinations of strings and tags for the level of the immediate constituents are built, which results in large sets of hypothetical constituent sequences.

For splits of unknown compounds, we presuppose that each immediate constituent should be found within the same close textual environment at least somewhere inside a large corpus. For derivatives, this holds only for hypothetical constituents which are free morphs or lexemes. In both cases, the sum of frequencies of the constituents in texts should be much lower for erroneous splits than the frequencies for correct segmentations.

Therefore, the free morphs and lexemes of these constituent sets are searched within their contexts. Here, we define contexts as the texts of a corpus in which the respective analyzed word form occurs. The 1.8 million articles of the annotated German Wikipedia Korpus of 2015 (Margaretha and Lüngen, 2014)[2] totals to 18.71 million word-form types. This provides a sample which is large enough for getting sufficiently many hits. 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 the tokenized and lemmatized forms. For each text, all frequencies of its lemmas and tokens are stored. For each text containing the input word form, the frequencies of its hypothetical constituents are retrieved.

For every word form W_{wf}, building and filtering the combinations of hypothetical immediate constituents produces a set of morphological splits. Each such morphological split $sp_{wf,s}$ consists of a sequence of hypothetical constituents $c_{wf,s,1}, c_{wf,s,2} \ldots c_{wf,s,n}$.

$$sp_{wf,s} = \{c_{wf,s,1}, c_{wf,s,2}, \ldots, c_{wf,s,n}\} \tag{1}$$

All texts comprising the word form W_{wf} are retrieved from the text indices. For each split and for every text T_t which contains the word form W_{wf}, the document frequencies ($df_1 \ldots df_m$) of the free hypothetical immediate constituents ($c_{wf,s,1} \ldots c_{wf,s,n}$) are being retrieved and summarized. This yields a text frequency score ($S_{wf,s,t}$) for each text and split of n constituents.

$$S_{wf,s,t} = \sum_{c=1}^{n} df_i \tag{2}$$

For every text, the highest text frequency score $Best_{wf,t}$ from all hypothetical analyses is chosen.

$$Best_{wf,t} = \max_{1,m} S_{wf,s,t} \tag{3}$$

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

Of all morphological analyses for W_{wf}, the one with the largest $Best_{wf,t}$ score is processed for the storage. (Equation 4).

$$BestSplitScore_{wf,t} = \max_{1,n} Best_{wf,t} \tag{4}$$

If no text contains W_{wf}, the score is 0. A missing hypothetical constituent inside a text containing W_{wf} leads to a document frequency of 0 for this constituent, which in principle can be compensated by the frequencies of the other constituents of the split sequence.

Bound morphemes are not considered for the BestSplitScores. As they are often ambiguous with function words, this avoids wrong segmentations such as (5-a) for *Abfertigung* 'check-in, dispatching' and leads to a preference to analyses as in (5-b) which then can be further expanded to the complex structure in (5-c).

(5) a. ♯Ab|Fertigung 'prefix, off|manufacture'
b. abfertigen|ung 'to check-in|suffix'
c. (*abfertigen* ab|(*fertigen* fertig|en))|ung

3.2.1 Morphological Segmentation based on Corpus Frequencies

If no text contains the word form W_{wf}, the corpus itself is considered as a textual environment in the widest sense. For example, the copulative adjective compound *preußisch-europäisch* 'Prussian-European' (6) is not in the Wikipedia text index, though its hypothetical constituents *preußisch* 'Prussian$_{adj}$', *Preuße* 'Prussian$_n$', *Europa* 'Europe' and *europäisch* 'European$_{adj}$' are.

(6)

preußisch-europäisch	P:preuße:<> isch - E:europa:ä isch	NN ADJSUFF HYPHEN NN ADJSUFF	<ADJ>
preußisch-europäisch	preußisch - E:europa:ä isch	ADJ HYPHEN NN ADJSUFF	<ADJ>

For all free morphemes and lexemes of each split, the sums of corpus frequencies are calculated. The hypothetical analysis with the highest value is chosen, and the morphological analysis with this score is processed for the storage. As a final back-off strategy, a default value of 0.1 is assigned to constituents which are unfound in the large corpus.

3.3 Reducing Sparse Data Effects

The derivation of hierarchical word structures from sequences of hypothetical constituents produces three types of error: These are: too few partitions, too many partitions, or a correct number but wrong subsets. In the first case, immediate constituents are less frequent than their constructions, in the second, hypothetical immediate constituents are more frequent than their constructions. The third error can be considered as a combination of the two previous ones.

3.3.1 Missing Constituents

The frequency of immediate constituents can be lower than the frequency of their word form. For example, the correct analysis for *Bambussieb* 'bamboo screen' is in (7-a). (7-b) shows the erroneous analysis of a compound as a monomorphemic lexeme. The overall text frequency of *Bambussieb* in the corpus is 3. None of these texts contains both constituents, and the frequencies of the compound outnumber the counts of the constituent *Sieb* 'sieve, screen' within these texts. This leads to a preference for the unsplit combination. However, this kind of analysis is desired or at least acceptable for real monomorphic and opaque word forms as in (7-c). Here, the SMOR split is false (7-c), however the look-up cannot obtain the word forms and hypothetical constituents of the erroneous analysis (7-d) within the same text, and the unsplit form is chosen.

(7) a. [[NN],[NN]] Bambus|Sieb
b. ♯[[NN, NN]] Bambussieb
c. ♯[[NN], [NN]] ♯da|rinnen '(there|trickle)'
d. [[NN, NN]] darinnen

Investigations on the lengths of German morphemes show that German simplex lexemes rarely possess more than 7 phonemes (98.41%) (Menzerath, 1954; Gerlach, 1982). The number of graphemes is proportional and slightly larger (Krott, 1996). Therefore, all word forms with more than 8 characters can be considered as candidates for polymorphemic analyses. For these, it is checked if a. BestScores were found only for splits comprising just one constituent but b. hypothetical splits with more than one constituent do exist. If these conditions hold, the word form undergoes a double check by the analysis based on the Wikipedia corpus frequencies. In Figure 2, this is indicated by the dashed line from the Recheck box.

3.3.2 Frequency Distributions of Constituents

The frequency-based weighting has a bias towards constructions with small constituents. While bound morphs which are often ambiguous with function words do not contribute to the scores (see 3.2), the problem is obvious for small frequent word forms as in (8).

(8) a. ♯Figur|Kombi|Nation 'figure|combi (short form of combination)|nation'
b. Figur|Kombination 'figure|combination'
c. Figur|(*Kombination* kombin|ation)

The relation between length and frequency of German morphs and lexemes was investigated by Köhler (1986) and Krott (2004). Both observed oscillating functions for morph and lexeme frequency depending on length. As the functional dependency is mutual and influenced by other factors such as the age of words and lexicon size, fitting typical distributions such as the mixed negative binomial distribution yields bad results and has no convincing linguistic interpretation. We found the same effects for the lexeme lengths and frequencies of our test corpus (see 4) and decided to use the frequencies of word length classes as inverse weights for the scores. For each constituent with a length of l characters, the frequency of its word length class L_l is used as an inverse proportional factor for the document frequencies (5). The weighted best scores are defined as in 3.2.

$$WeightedS_{wf,s,t} = \sum_{c=1}^{n} \frac{df_i}{freq(L_{l(c)})} \tag{5}$$

3.4 Substitution of Analyses

For all found best splits, the analyses for every immediate constituents are being searched in the databases and integrated into the analysis. Figure 2 shows an example for *Chefredakteurin* 'chief editor (female)'. The contextual search leads to the split *Chefredakteur|in* which can be refined by the analysis of *Chefredakteur* 'chief editor (male)'. These morphological splits are added to the new splits database.

4 Evaluation

The test data was extracted from *Korpus Magazin Lufthansa Bordbuch (MLD)* which is part of the DeReKo-2016-I (Institut für Deutsche Sprache, 2016) corpus.[3] We tokenized and lemmatized the texts by the TreeTagger (Schmid, 1999).[4] The resulting data comprises 276 texts with 260,114 tokens, 38,337 word-form types, and 27,902 lemma types. 15,622 of these lemma types are inside the databases of trees or monomorphemic words. This is a coverage of 55.99% with an accuracy of nearly 100% due to the quality of the CELEX and GermaNet data.

The remaining 44.01% of all lemma types were processed by SMOR and Moremorph with a coverage of 100%. We took the counts of the word length classes from the set of lemmatized tokens of the MLD corpus. Unknown words, e.g. numbers, are analyzed as simplex words. We took a sample of 1,006 word forms by extracting every 27th line of the produced output. For this evaluation, we distinguish a correct analysis for all levels, no or missing splits on the level of the ICs, flat or partially flat analyses, and erroneous segmentations.

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

Typical examples for analyses without any splits are word forms which are on a scale between adjectives and participle forms (9-a). Other forms occur more often than their constituents within a relatively small amount of contexts (9-b). Flat analyses are a typical outcome for words with sequences of short hypothetical immediate constituents (9-c). Sometimes, they are questionable but mostly close to a senseful interpretation on the surface morph-level. We found one erroneous split which was transferred from GermaNet, wrongly analyzed as endocentric compound instead of a syntagmatic construction (9-d). Besides this, wrong analyses are usually based on frequencies of homographs as in (9-e).

(9) a. folgend 'following', gewandt 'turned$_v$, skillful$_{adj}$'
b. Papierticket 'paper ticket', Metallkäfig 'metal cage', Tierärztin 'vet$_{fem}$'
c. Ab|Flug|Gate '(prefix, away|flight|Gate), departure gate', Roll|vor|Gang '♯?(to roll|prefix, before|gait), rolling procedure',
d. ♯?(Land|Nahme) '(land|"take"), settlement'
e. ♯(Parlament|(*arisch* ar|isch)) '(parliament|(*Aryan* Ar|ian), parliamentary'
f. rollen|(*Vorgang* (*vorgehen* vor|gehen)) 'to roll|(*procedure* (*to proceed* pro|ceed))'

The recheck analyses word forms which were otherwise annotated as consisting of one constituent. Therefore the chance for wrong complex analyses grows. For instance, due to a preference for small constituents, *Metallkäfig* receives the split ♯(Met|All|Käfig) '(mead|space|cage), metal cage'. However, the word-length weighting method changes this to the correct split, same as for *Rollvorgang* (9-c), (9-f). Table (9) presents a concise summary of the evaluation.

Table 1: Results of hybrid word analyzing

	correct analysis (all levels)	**no analysis**	**(partially) flat analysis**	**wrong analysis**
DBs	≈ 55.99%			
DBs + Context + Corpus Look-up	87.77%	7.45%	3.48%	1.29%
+ Recheck	92.44%	2.68%	2.88%	1.99%
+ Recheck + Weighting	93.34%	2.58%	2.78%	1.29%

For all procedures, we found 18 or less wrongly analyzed word forms inside the sample of a thousand words. This shows a good quality of the analysis. Preferences do neither exist for right-branching Schmid (2005) nor for left-branching, but rather for flat structures. 5,696 new entries were added to the monomorphemic lexemes and 8,448 to the new splits. Those analyses can be added to the knowledgment base with or without human supervision.

5 Summary

We presented a hybrid approach for deep-level morphological analysis. On the one hand, it is based on databases of previous work which we recycled and combined to a new form. On the other hand, flat structures from a morphological segmentation tool served as a starting point. All plausible combinations of the immediate constituents were evaluated by look-ups in textual environments of a large corpus or inside the set of all types as a back-off strategy. Biases towards small constituents with high frequencies on the one side and unsplit words on the other were tackled by insights from investigations in quantitative linguistics. The combination of the methods lead to an accuracy of 93% for complex structures and 98.7% for acceptable output.

6 Future Work

For improvement, there are two directions: using larger corpora, to possibly obtain a better fit of the wordlength-frequency relationship. On the other hand, inhomogeneous data can blur models. Therefore, analyzing words text by text could help to achieve larger contextual dependency and to find morphological structures fitting to the direct environment. This would result in different structures for orthographical words according to their contexts.

Acknowledgments

Work for this publication was partially supported by the German Research Foundation (DFG) under grant RU 1873/2-1. I especially thank Reinhard Rapp for the joint work, and Helmut Schmid for developing SMOR, for making it freely available, and for his cooperation and advice for making changes in the lexicons and transition rules.

References

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

Fabienne Cap. 2014. *Morphological processing of compounds for statistical machine translation*. Ph.D. thesis, Universität Stuttgart. https://doi.org/10.18419/opus-3474.

Stefanie Dipper. 2016. Tokenizer for German. https://www.linguistics.rub.de/~dipper/resources/tokenizer.html.

Rainer Gerlach. 1982. Zur Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie. In W. Lehfeldt and U. Strauss, editors, *Glottometrika 4*, Brockmeyer, Quantitative Linguistics 14, pages 95–102.

Alexander Geyken and Thomas Hanneforth. 2006. TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In *FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, Springer, Berlin/Heidelberg, volume 4002 of *LNCS*, pages 55–66. https://doi.org/10.1007/11780885_7.

Mariikka Haapalainen and Ari Majorin. 1995. GERTWOL und morphologische Disambiguierung für das Deutsche. http://www2.lingsoft.fi/doc/gercg/NODALIDA-poster.html.

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, Hissar, Bulgaria, 2011*. Association for Computational Linguistics, pages 420–426. http://www.aclweb.org/anthology/R11-1058.

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

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, April 12-17, 2003, Budapest, Hungary*. Association for Computational Linguistics, volume 1, pages 187–193. https://doi.org/10.3115/1067807.1067833.

Reinhard Köhler. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Quantitative Linguistics 31. Studienverlag Dr. N. Brockmeyer, Bochum.

Andrea Krott. 1996. Some remarks on the relation between word length and morpheme length. *Journal of Quantitative Linguistics* 3(1):29–37. https://doi.org/10.1080/09296179608590061.

Andrea Krott. 2004. Ein funktionalanalytisches Modell der Wortbildung [A functional analytical model of word formation]. In Reinhard Köhler, editor, *Korpuslinguistische Untersuchungen zur Quantitativen und Systemtheoretischen Linguistik [Corpus-linguistic Investigations of Quantitative and System-theoretical Linguistics]*, Elektronische Hochschulschriften an der Universität Trier, Trier, pages 75–126. http://ubt.opus.hbz-nrw.de/volltexte/2004/279/pdf/04_krott.pdf.

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 International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association (ELRA), pages 1848–1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Wolfgang Lezius. 1996. Morphologiesystem Morphy. In R. Hausser, editor, *Linguistische Verifikation. Dokumentation zur ersten Morpholympics 1994*. Niemeyer, Tübingen, pages 25–35. http://www.ims.uni-stuttgart.de/projekte/corplex/paper/lezius/molympic.pdf.

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, August 16, 2016, Berlin, Germany*. Association for Computational Linguistics, pages 76–81. https://doi.org/10.18653/v1/W16-2012.

Eliza Margaretha and Harald Lüngen. 2014. Building linguistic corpora from wikipedia articles and discussions. *Journal of Language Technology and Computational Linguistics. Special issue on building and annotating corpora of computer-mediated communication. Issues and challenges at the interface between computational and corpus linguistics* 29(2):59 – 82. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-33306, http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf.

Paul Menzerath. 1954. *Die Architektonik des deutschen Wortschatzes*. Phonetische Studien. Dümmler, Bonn ; Hannover ; Stuttgart.

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 Technologie, June 12-17, 2016, San Diego, California, USA*. Association for Computational Linguistics, pages 617–622. https://doi.org/10.18653/v1/N16-1075.

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*, Springer Netherlands, Dordrecht, pages 13–25. https://doi.org/10.1007/978-94-017-2390-9_2.

Helmut Schmid. 2005. Disambiguation of Morphological Structure using a PCFG. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 6-8 October, 2005, Vancouver, British Columbia, Canada*. Association for Computational Linguistics, pages 515–522. https://www.aclweb.org/anthology/H05-1065.

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 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association (ELRA). http://www.aclweb.org/anthology/L04-1275.

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, Prague, Czech Republic*. pages 146–160. https://aclweb.org/anthology/W17-7619.

Petra Steiner and Reinhard Rapp. in press. Building and Exploiting Lexical Databases for Morphological Parsing. In *Proceedings of The International Conference on Contemporary Issues in Data Science, March 5-8, 2019, Zanjan, Iran*. Springer, Lecture Notes in Computer Science.

Petra Steiner and Josef Ruppenhofer. 2018. Building a Morphological Treebank for German from a Linguistic Database. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan*. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1613.

Kyoko Sugisaki and Don Tuggener. 2018. German Compound Splitting Using the Compound Productivity of Morphemes. In *14th Conference on Natural Language Processing - KONVENS 2018*. Austrian Academy of Sciences Press, pages 141–147. https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/konvens18_16.pdf.

Marion Weller-Di Marco. 2017. Simple Compound Splitting for German. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), Valencia, Spain*. Association for Computational Linguistics, pages 161–166. https://doi.org/10.18653/v1/W17-1722.

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. http://aclweb.org/anthology/W/W13/W13-1807.pdf.

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, 11 August, 2016, Berlin, Germany*. Association for Computational Linguistics, pages 50–55. https://doi.org/10.18653/v1/W16-1807.

Patrick Ziering and Lonneke van der Plas. 2016. Towards Unsupervised and Language-independent Compound Splitting using Inflectional Morphological Transformations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, USA, June 12-17, 2016*. Association for Computational Linguistics, pages 644–653. https://www.aclweb.org/anthology/N16-1078.