Building a Morphological Treebank for German from a Linguistic Database

Petra Steiner, Josef Ruppenhofer
Institut für Deutsche Sprache
{ruppenhofer, steiner}@ids-mannheim.de

Abstract

German is a language with complex morphological processes. Its long and often ambiguous word forms present a bottleneck problem in natural language processing. As a step towards morphological analyses of high quality, this paper introduces a morphological treebank for German. It is derived from the linguistic database CELEX which is a standard resource for German morphology. We build on its refurbished, modernized and partially revised version. The derivation of the morphological trees is not trivial, especially for such cases of conversions which are morpho-semantically opaque and merely of diachronic interest. We develop solutions and present exemplary analyses. The resulting database comprises about 40,000 morphological trees of a German base vocabulary whose format and grade of detail can be chosen according to the requirements of the applications. The Perl scripts for the generation of the treebank are publicly available on github. In our discussion, we show some future directions for morphological treebanks. In particular, we aim at the combination with other reliable lexical resources such as GermaNet.

Keywords: treebank, morphology, word structure, deep-level morphological analyses, CELEX, German

1. Introduction

German is a language with complex processes of word formation, of which the most common are compounding, derivation and conversion. The resulting lexical units usually have long orthographical forms. Moreover, many word forms have more than one combinatorially possible analysis, as in Figure 1: *Hauptbahnhof* 'central station' consists of three morphs which can be combined in two ways on the level of immediate constituents but only the first combination is the correct structure.

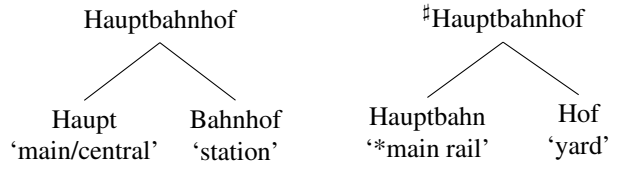

Figure 1: Ambiguous analysis of *Hauptbahnhof* 'central station'

Other word forms have ambiguous boundaries of morphs as in Figure 2 where the word form *Zugriff* 'grasp/access' is the product of a conversion process from *zugreifen* 'to grab/to grasp' and not a compound of other forms which could be erroneously recognized by a morphological analysis program.

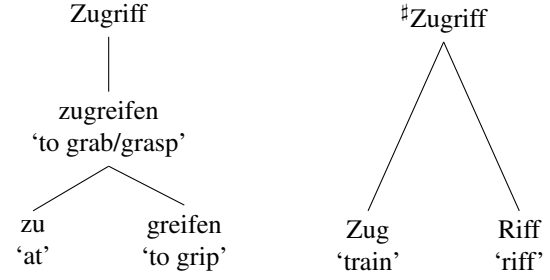

Figure 2: Ambiguous analysis of *Zugriff* 'grasp/access'

Morphological splitters for German such as Gertwol (Haapalainen and Majorin, 1995), MORPH (Hanrieder, 1996), SMOR (Schmid et al., 2004), or TAGH (Geyken and Hanneforth, 2006) generate many ambiguous analyses. Usually, this problem is approached by filtering procedures on the output analyses. For the ranking of the different morphological analyses, the geometric mean is a common score (Cap, 2014; Koehn and Knight, 2003; Steiner and Ruppenhofer, 2015). Another method is pattern matching with tokens (Henrich and Hinrichs, 2011) or comparisons of lemmas (Weller-Di Marco, 2017) or strings (Daiber et al., 2015) with corpus data. Ziering and van der Plas (2016) use normalization combined with ranking by the geometric mean. Würzner and Hanneforth (2013) apply a probabilistic context free grammar for parsing adjectives. Ma et al. (2016) apply Conditional Random Fields modeling for letter sequences. More recent approaches exploit semantic information for the ranking (Riedl and Biemann, 2016; Ziering et al., 2016). Besides Würzner and Hanneforth (2013), none of the above-mentioned authors tackled the challenge of generating deep-level analyses and with the exception of Henrich and Hinrichs (2011) who are using compounds from GermaNet (Hamp and Feldweg, 1997), all of them rely merely on corpus data as input for the heuristics and scores. However, carefully produced lexical data with morphological information would be a valuable asset for many applications in natural language processing. High quality morphologically deep-level analyses could especially be used as

1. input for statistical approaches for full morphologically parsing of German words
2. base of frequency counts for the testing of statistical hypotheses about morphological tendencies and laws
3. gold standards and test kits for morphological analyzers
4. morphological resources for morphological analyzers
5. input for textual analyses

Work on such kind of data is still in its beginning. This is shown in the following Section 2. where the related work is summarized in a concise way.
The work we present here is the generation of a morphological treebank for German. It is based on the German part of the refurbished CELEX database (Baayen et al., 1995), a manually constructed and human-supervised lexical resource. Section 3. describes this data with an emphasis on those parts which are relevant for the tree extraction process as well as the problems and flaws of the data. It also gives a sketch of the preprocessing. Section 4. presents the procedures we use. It starts with the extraction of all relevant information from the database, followed by the recursive construction of the morphological analyses. A heuristic for excluding unwanted diachronic information is presented, followed by details of the output format. The results of the script are presented in Section 5. The conclusion in Section 6. provides some further perspectives.

2. Related Work

Most German morphological data resources are restricted to lists of flat analyses. For instance, the test set of the 2009 workshop on statistical machine translation[1] was used by Cap (2014). It comprises 6,187 word tokens with splits on the upper level and interfixes removed. For example, in (1) the interfix *-s* and the hyphen have been deleted in the analysis.

(1) lexeme: *Abschreckungs-Ära* 'era of deterrence'
analysis: Abschreckung|Ära 'deterrence|era'

This is connected with some typical features of the much used morphological tool SMOR (Schmid et al., 2004). However, these interfixes are frequently marking boundaries between morphological constituents of higher levels. This is a reason why Steiner and Ruppenhofer (2015) modified the output of this tool to splits as (2).

(2) Abschreckung|s-|Ära

Henrich and Hinrichs (2011) augmented the GermaNet database with information on compound splits. This is restricted to nouns and does not provide interfixes or deep-level structures. However, in connection with this project, Steiner (2017)[2] derives deep-level structures with information on interfixes and grammatical properties from the GermaNet compounds which can be combined with analyses from CELEX.
DErivBase[3] (Zeller et al., 2013) comprises derivational families (word nests), however, the unsupervised generation of this derivational lexicon is based on heuristics of rules and string transformations. These rules do not always produce word families whose members are actually morphologically connected and the process of generation does not comply with linguistic evidence. However, the sets are produced as data for semantic (similarity) tasks and therefore do not claim grammatical correctness. Still, they contain some inconsistencies, e.g. the abridged word nest in (3) with its connection of formally similar words such as *Pause* 'pause, break' and *pausen* 'to calk, copy'. Zeller et al. (2014) assign evaluation measures to the lemma pairs of the nests for coping with this problem at least for the semantic level.

(3) $pausen_V$ 'to calk' – $abpausen_V$ 'to copy' – $pausieren_V$ 'to pause' – [...] –
$pausenlos_A$ 'without pause' – $Pause_{Nf}$ 'break' – $Zwischenpause_{Nf}$ 'short break'

The German part of the CELEX database (Baayen et al., 1995) comprises word tree information for a lexicon containing words of all parts of speech and is therefore an important source for deep-level morphological analyses of German, which are not available elsewhere. The linguistic information is combined with frequency information based on corpora (Burnage, 1995) which makes it useful for automated morphological analysis of unknown words. The original drawbacks of the German part of the database were an outdated format and use of obsolete orthographical conventions. However, these problems were tackled by Steiner (2016), so that the refurbished database yields a foundation for further exploitation. The lexicon with its 51,728 entries is relatively small but it covers a core vocabulary, similar to the small dictionary *Der kleine Wahrig* (Wahrig-Burfeind and Bertelsmann, 2007).
Shafaei et al. (2017) use the German data of CELEX for inferring derivational families which are more precise than DErivBase. The produced database DErivCELEX is drawn from the original CELEX version with its old orthographical standard[4] and therefore contains some inconsistencies and mistakes from string transformations such as (4). As some derivations of CELEX include diachronic information which became intransparent, the word nests might contain some word forms whose relatedness is rather historical than semantic, e.g. in the abridged set in (5) where constituents of *Flüssigkeit* 'fluid, liquid', *Floß* 'raft', *überflüssig* 'superfluous' and *beeinflussen* 'to influence' are all diachronically linked to *Fluss* 'river' (which is missing in DErivCELEX) and *fließen* 'to flow'.

(4) $^*blaün_V$ for *bläuen* 'to blue'

(5) $durchfließen_V$ 'to flow through' – $Floß_N$ 'raft' – $überflüssig_A$ 'superfluous' – $Zufluß_N$ 'feeder' – $unbeeinflußbar_A$ 'uninfluenceable' – [...] – $flößbar_A$ 'floatable' – $zusammenfließen_V$ 'to flow together' – $Beeinflussung_N$ 'influence' – $Zusammenfluß_N$ 'confluence' – $Flüssigkeit_N$ 'fluid, liquid' – [...] – $fließen_V$ 'to flow' – [...] – $beeinflussen_V$ 'to influence' – $beeinflußbar_A$ 'influenceable'

Just like DErivBase, DErivCELEX does not contain morphological analyses, but word family sets. DErivCELEX inherits the quality of CELEX with its manually corrected analyses; therefore it does not exhibit errors such as in (3). Shafaei et al. (2017) assert that CELEX does not treat prefixation as a form of derivation. In general, this assertion is

[1] http://www.statmt.org/wmt09/translation-task.html
[2] see https://github.com/petrasteiner/morphology
[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/DErivBase/DErivBase-v2.0.zip
[4] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/DErivBase/DErivCelex-v1.txt

unjustified, though some first constituents of verbs are classified as free morphs which Shafaei et al. (2017) consider as prefixes. The CELEX classification is justifiable from a linguistic viewpoint concerning the differences between prefixes and particles. However, as this restricts the sets of the derivational families, Shafaei et al. (2017) produce a second database based on a wider definition.
Dutch morphological analysis is covered by CELEX too. However, we are not aware of any exploitation for morphological deep-level analyses. For English, Cotterell et al. (2016) reanalyse a part of CELEX deep-level morphological analyses and thus generate 7,454 morphological parses. For other languages, there are some resources of derivational families such as in CroDeriV for Croatian (Filko and Šojat, 2017), Démonette for French (Hathout and Namer, 2016), DeriNet for Czech (Žabokrtský et al., 2016) or DerIvaTario for Italian (Talamo et al., 2016). These could be exploited for the derivation of morphological trees. However, automatic analyses are not trivial if generation rules are incomplete or multiple derivational rule paths are possible. Besides this, compounds are not considered by these lists.

3. The Refurbished CELEX-German Database

Developed in the early Nineties, the original CELEX database coding comprised a workaround for special characters. In German, these are mainly umlauts and signs such as *ß*. Furthermore, it uses an out-dated spelling convention which makes the lexicon partially incompatible with text written after 1996. For instance, the modern spelling of the original CELEX entry *Einfluß* 'influence' is *Einfluss*.
In Steiner (2016) entries such as for the lemma *Einflussbereich* 'range of influence' (6) for the orthographical part of the database and (7) for the morphologically database were aligned as in (8). Please note that these examples only present the essential and abridged information of the structure information and the morphological trees. A database with modern encoding but old spelling with *ß* was also derived as in (9). Trees as in Figure (3) could be derived directly from the database.

(6) 10236\Einflu$bereich\8\Ein-flu$-be-reich\N

(7) 10236\Einflussbereich\Einfluss+Bereich\NN\
((((ein)[V|.V],(fliess)[V])[V])[N],
((be)[N|.N],(Reich)[N])[N])[N]

(8) Einflussbereich\Einfluss+Bereich\NN\
((((ein)[V|.V],(fließ)[V])[V])[N],
((be)[N|.N],(Reich)[N])[N])[N]

(9) Einflußbereich\Einfluß+Bereich\NN\
((((ein)[V|.V],(fließ)[V])[V])[N],
((be)[N|.N],(Reich)[N])[N])[N]

However, trees of this kind have some gaps: they do not contain categorial information for affixes nor for the derivation process, e.g. the noun *Einfluss* 'influence'. Therefore, simple transformations of the data would yield only incomplete derivations.

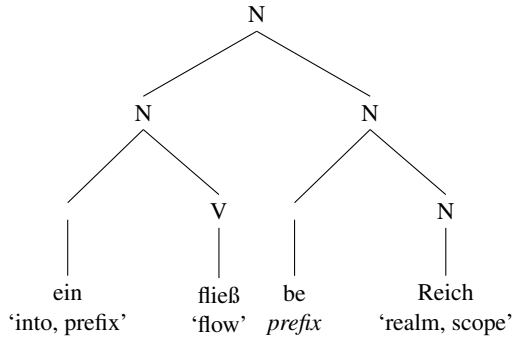

Figure 3: Rudimentary morphological analysis of *Einflussbereich* 'range of influence'

Another drawback is the missing information on the infinitive stem. While in (8) this would be *-en* for *fließen*, in the analysis of *Abenddämmerung* 'evening dawn = nightfall' in (10) there is an elision of the schwa of the infinitive stem *dämmern* 'to dawn'.
Some derivations in the German CELEX database provide diachronic information which is correct but often unwanted for many applications, for example in (11) (*Schnellzug* 'fast(-speed) train') where *Zug* 'train' is diachronically derived from *ziehen* 'to draw', see Figure 4. This analysis is completely opaque from a synchronic point of view. On the other hand, some derivations such as the ablaut changes between *fließen* and *Einfluss* in Figure 3 or *gehen* 'to go' and *Gang* 'gait,path,aisle' in *Abgangszeugnis* 'leaving certificate' (12) could be of interest.

(10) 111\Abenddämmerung\Abend+Dämmerung\NN
((Abend)[N],((dämmer)[V],(ung)[N|V.])[N])[N]

(11) 34419\Schnellzug \schnell+Zug\AN\
((schnell)[A],((zieh)[V])[N])[N]

(12) 207\Abgangszeugnis\Abgang+s+Zeugnis\NxN\
((((ab)[V|.V],(geh)[V])[V])[N],(s)[N|N.N],
((zeug)[V], (nis)[N|V.])[N])[N]

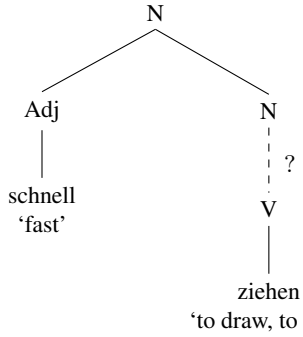

Figure 4: Rudimentary and questionable morphological analysis of *Schnellzug* 'fast train'

Interfixes can be inferred from the database entry. In Figure 5, the interfix is represented as an affix (*x*) within the categories of the immediate constituent structure. In the

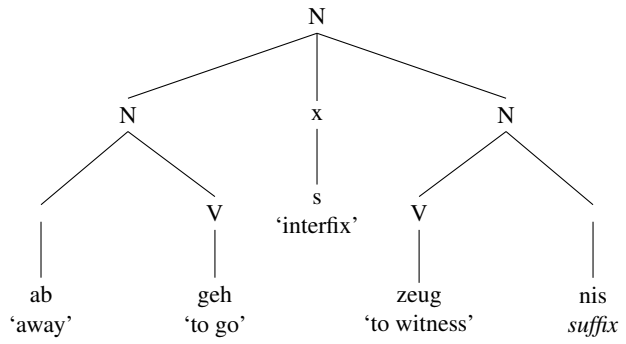

Figure 5: Rudimentary morphological analysis of *Abgangszeugnis* ‘leaving certificate’

CELEX entries they are part of the categorial description of the immediate structures, such as *NxN* within the example (12). As every complex entry has such information on the immediate constituents and their categories, it is possible to collect this information recursively and top-down from the CELEX entries.

Though most of its data are flawless, the original CELEX database contains some mistakes which were not treated by the refurbishment of Steiner (2016) which covered only changes of coding and spelling. We found

- missing constituents and missing part-of-speech information within the morphological trees
- missing constituents within the field of immediate constituency information
- inconsistent morphological analyses, such as *Kenntnisnahme* ‘notice, attention’ in Figure 6 which should have been analysed as a conversion (Zusammenrückung), similar to *Maßnahme* ‘measure, step’ which in CELEX is analyzed as resulting from a conversion of *maßnehmen* ‘to take measures’.

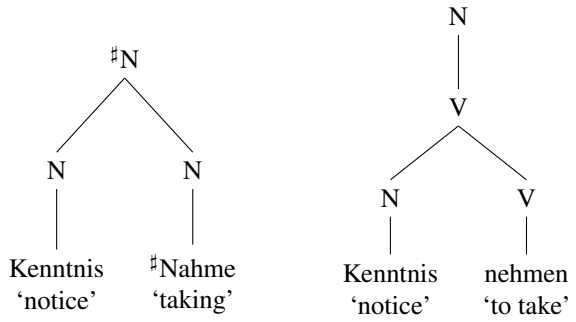

Figure 6: Erroneous and correct morphological analysis of *Kenntnisnahme* ‘notice’

We augmented the script for the transformation to a modern standard by 18 additional rules, which covered 65 instances before we could use the data for extracting the morphological trees. We are aware of the fact that we could not find all mistakes. The Perl script *OrthCELEX.pl* for the refurbishment and correction of the German CELEX data is available on github.[5]

4. Procedures

The extraction of the CELEX-German treebank is based on the refurbished and corrected database which we have described in the last section. Figure 7 shows the dataflow and the main procedures.

We do not produce one single treebank, but leave it to the users which format and information they choose for the trees they intend to build. For example, semantic word nests might require less diachronic information than finding anaphora in texts. Conversions can be of interest or not. The generating script provides some parameters for refinements and output formats. We first extract all the information which could be required and then build the trees recursively and top-down according to the options.

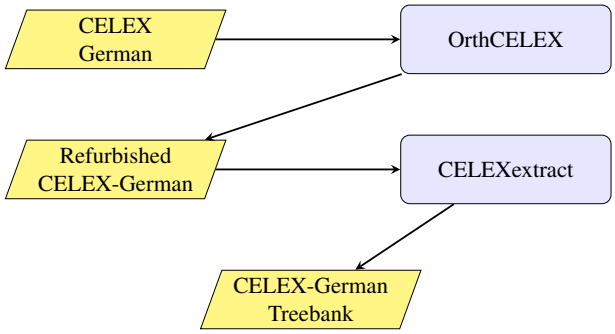

Figure 7: The dataflow from CELEX-German to the CELEX Treebank

4.1. Data Extraction

We start with extracting all relevant information. Some forms can be assigned more than one part of speech as in (13), or more than one gender as in 14, or they are morphologically ambiguous as in (15) and Figure 8. Therefore we build an inverted index of all lemmas.

(13) a. aber ‘but, conj’
b. aber ‘really, intensifier’

(14) a. Band ‘volume/book, noun’
b. Band ‘band (music), noun’
c. Band ‘ribbon/strap, noun’

(15) a. erzen ‘made out of ore, bronze, adj’
b. erzen ‘to address by *er*, verb’

We extract all immediate constituents and also their categories, then we internally add the infinitive forms of the verbs which are included within these entries. This is necessary for finding these forms within the inverted index of the entries. Also we refurbish the German syntactic database of CELEX to the modern standard and extract the parts of speech of the entries.

As the users can choose if they would like to generate not just compounds and derivatives but also conversions, we extract the relevant information for this word-formation class too.

[5] see https://github.com/petrasteiner/morphology

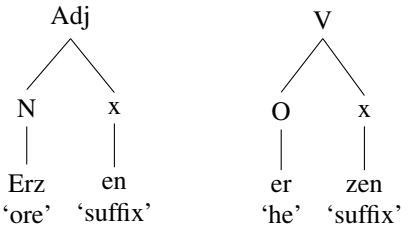

Figure 8: The ambiguity of the form *erzen*

4.2. Building the Trees

For each entry of the morphological database, the procedure starts from the list of its immediate constituents and recursively collects all information from the entries of the constituents. Algorithm 1 presents the recursive process. Table 1 shows the parameters.

4.3. Prevention of Diachronic Information

Diachronic information, as in example (11) and Figure 4 with *Zug* ‘train’ being diachronically derived from *ziehen* ‘to draw’, can be of interest, however, for many applications it is considered as unnecessary or even disturbing. Therefore, the script permits users to choose a threshold of similarity within the range of [0:1] which is compared to a measure using the Levenshtein distance.

For accepting or rejecting two parts of words as morphologically related, the procedure will cut two forms f_1, f_2 with length l_1 and l_2 to the strings s_1, s_2 of the smaller length $(min(l_1, l_2))$ and calculate the Levensthein distance (LD) of these. Special characters such as *ä* or *ß* are transformed to *a* and *ss*, uppercase characters to lowercase. Then the quotient of both values is compared to a threshold *t* as in (16):

$$\frac{LD(s_1, s_2)}{min(l_1, l_2)} < t \tag{16}$$

For example, in (17) both the derived form (e.g. $f_1 = Zug$) and its component (e.g. $f_2 = zieh$) are cut to the smaller size of these forms in lowercase letters. In this case, that yields the strings $s_1 = zug$ and $s_2 = zie$. After this, the quotient of $LD(s_1, s_2)$ and the smaller length is compared to the threshold. (17) shows that the analysis for this case would be interrupted for a threshold below $0.\overline{6}$. A value of 1 would show total dissimilarity, one of 0 absolute similarity.

$$\frac{LD(zug, zie)}{min(3, 4)} = \frac{2}{3} \tag{17}$$

In case that singular variations were needed, we also added a small list of exceptions.

4.4. Output Formats

Our tool supports various output formats. Table 1 lists the optional parameters which are available. The depth of the morphological trees can be determined, same as including the analysis of conversions or which linguistic information should be provided, e.g. the parts of speech and classes of the bound morphs. If the threshold for the Levenshtein-based measure is defined, the top-down generation of the morphological trees will be stopped for elements which are more dissimilar to each other than permitted. For the output style, the user can choose parentheses or a notation with pipe bars (”|”) for the splits on the same level.

Algorithm 1: Building the morphological treebank

```
Input: CELEX-German revised
Output: A Morphological Treebank
initialization of parameters: depths of analysis,
  levenshtein threshold, linguistic information,
  parts of speech, style of output;
forall entries of CELEX do
    if entry is complex or a conversion then
        foreach constituent of entry do
            if constituent is simplex
              or depth of analysis reached then
                retrieve linguistic information/PoS
                as required;
                return linguistic information and
                  constituent
            end
            else
                foreach part of constituent do
                    depth of analysis++;
                    analysedeeper part with
                      parameters and depth;
                    return result of analysedeeper
                end
            end
        end
    end
end

 sub analysedeeper part (parameters and level)
  if part is simplex
  or depth of analysis reached
  then
    retrieve linguistic information/PoS as required;
    return linguistic information and part
end
else
    foreach subpart of part do
        analysedeeper subpart
          if levenshtein threshold and
          analysedeeper subpart is dissimilar then
            skip deeper analysis;
            retrieve linguistic information/PoS as
              required;
            return subpart
        end
        else
            return result of analysedeeper subpart
        end
    end
end
```

	Parameters
•	Depth of analysis for compounds and derivatives
•	Analysis of conversions
•	Depth of analysis for conversions
•	Lingustic information of the morphs
•	Threshold for the Levenshtein measure
•	Style of format

Table 1: Parameters for the tree generation

5. Results

The list of all word-formation products of the German database (compounds, derivatives, results of conversions) comprises 40,097 entries.

5.1. Coverage

We tested the coverage of this treebank on the *Korpus Magazin Lufthansa Bordbuch (MLD)* which is part of the DeReKo-2016-I (Institut für Deutsche Sprache, 2016) corpus[6]. 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 xml-annotated data comprises 276 texts with 5,202 paragraphs, 16,046 sentences and 260,115 tokens. The number of word-form types is 38,337. Of these types, 5,435 are included in the CELEX-derived treebank. If we add all entries, including also the simplex forms, the overlap of the types is 8,622. We are comparing a list of lemmas with a list of word forms, this means that not every full form can be covered. Therefore, the overlap is a good start, especially as (longer) word-formation products could be analyzed in combination with a word splitter for flat structures.

5.2. Output

The following shows the entries of *Einflussbereich*, *Schnellzug*, and *Abgangszeugnis*. For the parameter setting of all linguistic information, the notation with |, and a Levenshtein threshold of 0.6, the results are presented in (18), for parenthesis notation and no restrictions on diachronic conversions in (19) and for a flat representation of the immediate constituents see (20).

(18)
```
Einflussbereich
(*Einfluss_N*
  (*einfließen_V*
    ein_x|
    fließen_V))|
(*Bereich_N*
  be_x|
  Reich_N)

Schnellzug
 schnell_A|
 (Zug_N)

Abgangszeugnis
(*Abgang_N*
   (*abgehen_V*
     ab_x|
     gehen_V))|
 s_x|
 (*Zeugnis_N*
   zeugen_V|
   nis_x)
```

(19)
```
Einflussbereich
 (*Einfluss_N*
  (*einfließen_V*
    (ein_x)
    (fließen_V)))
 (*Bereich_N*
   (be_x)
   (Reich_N))

Schnellzug
 (schnell_A)
 (*Zug_N*
    ziehen_V)

Abgangszeugnis
 (*Abgang_N*
   (*abgehen_V*
    (ab_x)
    (gehen_V)))
 (s_x)
 (*Zeugnis_N*
   (zeugen_V)
   (nis_x))
```

(20)
```
Einflussbereich
 Einfluss_N|
 Bereich_N

Schnellzug
 schnell_A|
 Zug_N

Abgangszeugnis
 Abgang_N|
 s_x|
 Zeugnis_N
```

Figures 9 and 10 show the complete analyses for *Einflussbereich* and *Abgangszeugnis* with all intermediate constituents.

6. Conclusion and Further Perspectives

This article introduces to the first German morphological treebank. Its form and output can be determined by the user of the Perl script *CELEXextract.pl* which is available on our repository.[7]

The possible analyses comprise compounds, derivatives and conversions of different depths and linguistic information as is required. The current database is relatively small,

[6] see Kupietz et al. (2010) for further information

[7] see https://github.com/petrasteiner/morphology

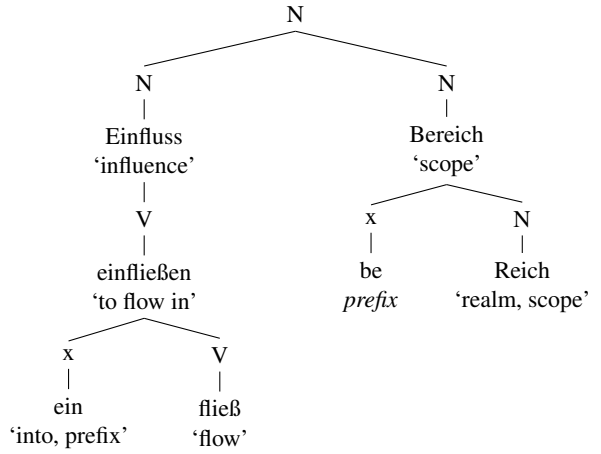

Figure 9: Complete morphological analysis of *Einflussbereich* ‘range of influence’

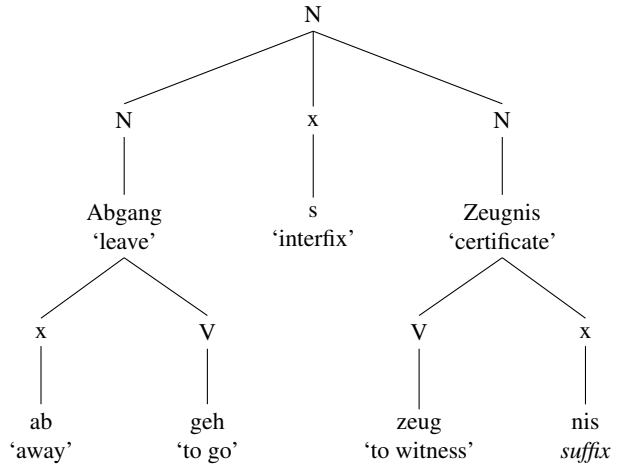

Figure 10: Complete morphological analysis of *Abgangszeugnis* ‘leaving certificate’

however, it will be augmented by other sources. This work has already started. Recently, Steiner (2017) combined the splits of the nominal compounds of GermaNet (Henrich and Hinrichs, 2011) with the more fine-grained analyses of CELEX’s basic vocabulary. While the GermaNet compounds on their own yield about 68,000 trees, here the recursive production of the morphological trees stops as soon as derivatives are reached. But merging both resources results in a German Treebank of ca. 100,000 analyses comprising the processes of compounding and derivation for each entry.

Compared to automatically inducing morphological resources, which then have to be cleaned and/or evaluated, the effort of using manually produced data for the induction of deep morphological analyses is relatively small and the effect is rewarding.

On the foundation of the existing database, more complex words can be analyzed in combination with a morphological splitter for flat structures. This method enlarges the coverage by the combinatorial potential of language and will avoid the abundance of ambiguous word analyses.

7. Acknowledgements

The author was supported by the German Research Foundation (DFG) under grant RU 1873/2-1. We would like to thank the reviewers for their valuable feedback.

8. Bibliographical References

Burnage, G. (1995). CELEX: A Guide for Users. In Harald Baayen, et al., editors, *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, Philadelphia, PA.

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

Cotterell, R., Kumar, A., and Schütze, H. (2016). Morphological segmentation inside-out language processing. In Jian Su, et al., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2325–2330. The Association for Computational Linguistics.

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. ÚFAL MFF UK.

Dipper, S. (2016). Tokenizer for German.

Filko, M. and Šojat, K. (2017). Expansion of the derivational database for Croatian. In *First Workshop on Resources and Tools for Derivational Morphology (DeriMo)*.

Geyken, A. and Hanneforth, T. (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.

Haapalainen, M. and Majorin, A. (1995). Gertwol und morphologische Disambiguierung für das Deutsche. In *Proceedings of the 10th Nordic Conference on Computational Linguistics, Helsinki, Finland*.

Hamp, B. and Feldweg, H. (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.

Hanrieder, G. (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.

Hathout, N. and Namer, F. (2016). Giving lexical resources a second life: Démonette, a multi-sourced morpho-semantic network for French. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1084–1091, Paris, France. European Language Resources Association (ELRA).

Henrich, V. and Hinrichs, E. (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.

Koehn, P. and Knight, K. (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.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German reference corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari, et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1848–1854, Valletta, Malta. European Language Resources Association (ELRA).

Ma, J., Henrich, V., and Hinrichs, E. (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.

Riedl, M. and Biemann, C. (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.

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)*. European Language Resources Association (ELRA).

Shafaei, E., Frassinelli, D., Lapesa, G., and Padó, S. (2017). DErivCELEX: Development and evaluation of a German derivational morphology lexicon based on CELEX. In *Proceedings of the DeriMo workshop*, Milan, Italy.

Steiner, P. and Ruppenhofer, J. (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 October2015*, pages 49–57.

Steiner, P. (2016). Refurbishing a morphological database for German. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*, pages 1103–1108. European Language Resources Association (ELRA).

Steiner, P. (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.

Talamo, L., Celata, C., and Bertinetto, P. M. (2016). Derivatario: An annotated lexicon of italian derivatives. *Word Structure*, 9(1):72–102.

Wahrig-Burfeind, R. and Bertelsmann, G. L. (2007). *Der kleine Wahrig: Wörterbuch der deutschen Sprache [der deutsche Grundwortschatz in mehr als 25000 Stichwörtern und 120000 Anwendungsbeispielen; mit umfassenden Informationen zur Wortbedeutung und detaillierten Angaben zu grammatischen und orthografischen Aspekten der deutschen Gegenwartssprache]*. Wissen Media Verlag.

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. Association for Computational Linguistics.

Würzner, K.-M. and Hanneforth, T. (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.

Žabokrtský, Z., Sevcikova, M., Straka, M., Vidra, J., and Limburská, A. (2016). Merging data resources for inflectional and derivational morphology in Czech. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Zeller, B., Šnajder, J., and Padó, S. (2013). DerivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1201–1211. Association for Computational Linguistics.

Zeller, B., Padó, S., and Šnajder, J. (2014). Towards semantic validation of a derivational lexicon. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1728–1739, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Ziering, P. and van der Plas, L. (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.

Ziering, P., Müller, S., and van der Plas, L. (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, August. Association for Computational Linguistics.

9. Language Resource References

Baayen, Harald and Piepenbrock, Richard and Gulikers, Léon. (1995). *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, 1.0, ISLRN 204-698-863-053-1.

Institut für Deutsche Sprache. (2016). *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2016-I (Release from 31.03.2016)*. Institut für Deutsche Sprache, 1.0.