Merging the Trees
Building a Morphological Treebank for German from Two Resources

Petra Steiner
Institut für Deutsche Sprache
steiner@ids-mannheim.de

Abstract

This paper deals with the creation of the first morphological treebank for German by merging two pre-existing linguistic databases. The first of these is the linguistic database *CELEX* which is a standard resource for German morphology. We build on its refurbished and modernized version. The second resource is *GermaNet*, a lexical-semantic network which also provides partial markup for compounds. We describe the state of the art and the essential characteristics of both databases and our latest revisions. As the merging involves two data sources with distinct annotation schemes, the derivation of the morphological trees for the unified resource is not trivial. We discuss how we overcome problems with the data and format, in particular how we deal with overlaps and complementary scopes. The resulting database comprises about 100,000 trees whose format can be chosen according to the requirements of the application at hand. In our discussion, we show some future directions for morphological treebanks. The Perl script for the generation of the data from the sources will be made publicly available on our website.

1 Introduction

Lexical productivity is a characteristic for German word formation. This leads to bottleneck problems in different fields such as the building of terminology or Information Retrieval. Concerning the morphological analyses and structures, there are three main problems:

A. the wealth of ambiguous forms on the level of morph segmentation
B. the lack of deeper structural analyses in current approaches
C. for morphological analysis in general, the lack of frequency counts or a robust estimation for affixes.

A morphological treebank of the most common lemmas or word forms of German can serve as a starting point for addressing all of these issues. Although the demand for such a morphological treebank with hierarchical analyses was recognized some time ago (Zielinski and Simon, 2009, 230), to our knowledge, morphological treebanks for German do not exist so far, besides some mostly internally used gold standards. Deep morphological analyses can be used as

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

We derive a morphological treebank for German from two different databases: the first resource is the linguistic database *CELEX* which is a standard resource for German morphology. The second resource is the *GermaNet* database which contains partial markup for compounds.

Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 146–160,
Prague, Czech Republic, January 23–24, 2018.

Keywords: morphological analysis, morphological treebank, German, GermaNet, compounding

Section 2 describes the current state of research for German deep-level morphological data. The first part of Section 3 describes the German part of the refurbished CELEX database with an emphasis on the data which are relevant for the tree extraction process as well as problems and errors in the data. It also gives a sketch of the preprocessing. The second part deals with the GermaNet (GN) database and the characteristics that are relevant for our project. Section 4 presents the procedures we use. It starts with the extraction of all relevant information from both databases, followed by the recursive construction of the morphological analyses. The derivation of the morphological trees for both sources is not trivial and we show how we overcome problems with the data and format. In Section 5, we show how we merge the two sources which have distinct annotation styles as well as overlaps and complementary scopes in their morphological classifications. We discuss the decisions used for the classification underlying our unified annotation. The results of the script are presented in Section 6. The resulting database comprises about 100,000 morphological trees whose format can be chosen according to the requirements of the applications. The conclusion in Section 7 provides some future directions for morphological treebanks. The Perl script for the generation of the data from the sources will be made publicly available on our website.

2 Related work

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. Therefore, so far the main concern of morphological analysers for German is finding the correct splits on the level of the morphs. Morphological segmentation tools for German such as SMOR (Schmid et al., 2004), Gertwol (Haapalainen and Majorin, 1995), MORPH (Hanrieder, 1996), TAGH (Geyken and Hanneforth, 2006) generate dozens of analyses for relatively simple words. For instance, *Kellerassel* "common rough woodlouse" could be erroneously segmented to ♯*Kelle*|*Rassel* "(ladle|rattle)" instead of *Keller*|*Assel* (basement|woodlouse) common rough woodlouse". Also, there are many sets of homonyms comprising both free and bound morphemes. For example, the form *bar* is a suffix in *machbar* mach|bar (make|able) "feasible", a free morph in *Hotelbar* (*hotel*|*bar*) "hotel bar" and a sequence without synchronically transparent meaning in *Nachbardistrikt* "neighboring district" which can be wrongly analysed to ♯*nach*|*Bar*|*Distrikt* (after|bar|district) (see Figure 1).

Figure 1: Ambiguous analysis of *Nachbardistrikt*

This ambiguity problem has been tackled by using ranking scores for the different morphological analyses. For example, Cap (2014) and Koehn and Knight (2003) use the geometric mean as a weighting measure for each possible analyses of SMOR and then choose the one with the highest rank. Another possibility are methods of exploiting the sequence of letters, e.g by pattern matching with tokens (Henrich and Hinrichs, 2011, 422), lemmas (Weller-Di Marco, 2017), or normalization (Ziering and van der Plas, 2016) which is 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.

More recent approaches exploit semantic information for the ranking. Riedl and Biemann (2016) take sets of constituent candidates they generate by combining a compound splitter and look-ups of similar terms inside a distributional thesaurus generated from a large corpus. 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 results of former splits, which were produced by morphological segmentation tools such as SMOR. The re-ranking shows a slight improvement over the initial values, while the pure distributional similarities were inferior to the initial results from the splitter. The reason for this is mainly the rate of word ambiguity which for large corpora is mirrored within the distributional patterns.

Most tools for word analyses of German word forms provide flat sequences of morphs or morphemes but no hierarchical parses which could give important information for word sense disambiguation. Only Würzner and Hanneforth (2013) tackle the problem of full morphological parsing, restricted to adjectives, by using a probabilistic context free grammar for parsing. Steiner and Ruppenhofer (2015) developed a method for building parts of morphological structures by reducing the set of all possible low-level combinations by ranking SMOR splits with the gmean score. They derived the frequencies from different lexical and textual sources, showing some effects which hint at the importance of carefully choosing the source of frequency counts.

Ziering et al. (2016) discuss left-branching compounds consisting of three lexemes such as *Arbeitsplatz-mangel* "(Arbeit|Platz|Mangel) (work|place|lack) job scarcity". Their distributional semantic modelling fails to find the correct binary split, if the head (here *Mangel* "lack") is too ambiguous to correlate strongly with the first part (here *Arbeitsplatz* "employment"). 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. Their evaluation however is based only on the binary compounds of GermaNet (Hamp and Feldweg, 1997).

All these approaches build strongly upon corpus data but none of them uses lexical data. Only Henrich and Hinrichs (2011) enrich the output of morphological segmentation with information from GermaNet to disambiguate such structures. This can yield hierarchical structures but presupposes that the entries for the components exist inside the database.

Databases of correct morphological splits and deep-level analyses could save a lot of effort, as there are almost no cases of forms with two different analyses which are really used, even if structure and splits can be analysed ambiguously. The second analysis in Figure (1) will hardly ever occur in real text. At most, it could be merely understood as a pun.

In most cases, 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] which was used by Cap (2014). It comprises 6,187 word tokens with binary top-level splits. Henrich and Hinrichs (2011) augmented the GermaNet database with information on noun compound splits of the top-level. DErivBase (Zeller et al., 2013) comprises derivational families (word nests) and could be used to infer derivational trees from its sets and rules, however, it is based on heuristics and therefore contains some errors.

The only publicly available source which comprises German word tree information is the German part of the CELEX database (Baayen et al., 1995). The linguistic information is combined with frequency information based on corpora (Burnage, 1995) which makes it useful for automated morphological analysis of unknown words.

That CELEX is a standard resource for research in morphology is demonstrated by Shafaei et al. (2017) who use its German data for inferring derivational families (DErivCELEX) which are more precise than DErivBase. This data is obviously drawn from the original CELEX version with its old orthographical standard.[2] Shafaei et al. (2017) claim that CELEX does not treat prefixation as a form of derivation. In general, this assertion is unjustified, though some first constituents of verbs are classified as free morphs which Shafaei et al. (2017) consider as prefixes. While the CELEX classification is justifiable from a linguistic viewpoint of consistency and difference between prefixes and particles, this proves as an error source for the algorithms of derivational families. For this reason, a second version of DErivCELEX is based on some "pragmatic changes" in categorization concerning compound verbs.

[1] http://www.statmt.org/wmt09/translation-task.html

[2] cf. http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/DErivBase/DErivCelex-v1.txt

Cotterell et al. (2016) reanalyse part of the deep-level morphological analyses for English and thus generate 7,454 morphological parses which to our knowledge is the only morphological treebank for English besides the aforementioned. Dutch morphological analysis is covered by CELEX too. For other languages, the situation is even less fortunate. But if there are resources of derivational families with information on their generating rules such as in CroDeriV (Filko and Šojat, 2017) for Croatian, Démonette for French (Hathout and Namer, 2016), DeriNet for Czech (Žabokrtský et al., 2016) or DerIvaTario for Italian (Talamo et al., 2016), hierarchical trees could be derived though compounds are not considered by these lists.

The original drawbacks of the German part of the CELEX database were an outdated format and use of former orthographical conventions. However, these problems were tackled by Steiner (2016), and so the database yields a foundation for further exploitation. We decided to take it as the foundation for the morphological treebank and then augment it by other sources, the first of which is the GermaNet database.

3 Lexical resources for morphological trees

3.1 The Refurbished CELEX-German Database

The CELEX database comprises 51,728 entries of which 38,650 are derivates or compounds and 2,402 conversions. This seems to be a small set, however, the lemmas are similar to the small dictionary *Der kleine Wahrig* (Wahrig-Burfeind and Bertelsmann, 2007) which represents the core vocabulary for German. Being developed in the early Nineties, the original CELEX database coding comprised a workaround for special characters. In German, these are mainly umlauts and characters 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 *Abschluß* 'conclusion' is *Abschluss*. About 20 percent of the data is in an outdated format. Steiner (2016) refurbished the encoding and the spelling of the database completely. A version with modern encoding but old spelling was also created. Now, trees as in Figure (2) and (3) can be derived from the database.

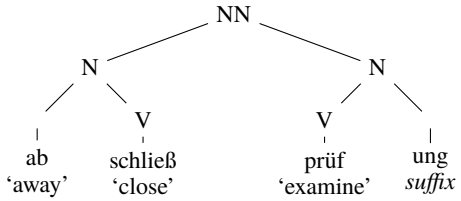

Figure 2: Morphological analysis of *Abschlussprüfung* 'final exam'

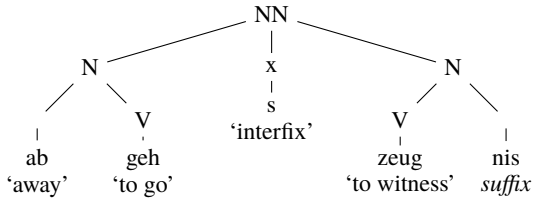

Figure 3: Morphological analysis of *Abgangszeugnis* 'leaving certificate'

However, these kinds of trees do not contain categorial information for affixes nor for the derivation process, e.g. the noun *Abschluss* 'finalization' in the derivation of (2). Morever, some derivations in the German CELEX database provide diachronic information which is correct but often unwanted for many applications, for example in *Abdrift* 'leeway' in example (1) which is diachronically derived from *treiben* 'to float'. On the other hand, some derivations such as the ablaut change between *gehen* 'to go' and *Gang* 'gait,path,aisle' in *Abgangszeugnis* 'leaving certificate' in example (2) could be of interest.[3]

(1) 97\Abdrift \ab+drift\xV\. . . \((ab)[N|.V],((treib)[V])[V])[N]

(2) 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 3 shows that the filler letters (interfix)[4] can be inferred from the database entry, where they are

[3]Please note that these examples of CELEX entries only present the essential and abridged information of the structure information and the morphological trees.

[4]Depending upon the framework, these entities are also called *Fugenmorpheme*. However, their morphological and phonological status can be discussed, and we prefer the term *filler letters* which refers to the form.

represented within the categories of the immediate constituent structure. As every complex entry has this information, this enables one to recursively collect them from the entries.

Though most of its data is free of errors, the original CELEX database contains some mistakes which were not treated by the refurbishment of Steiner (2016) which involved only changes of coding and spelling. We found missing constituents and missing part of speech information within the morphological trees and within the field of immediate constituency information as well as inconsistent morphological analyses. 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.

3.2 Compound Analyses from GermaNet

Henrich and Hinrichs (2011) augmented the GermaNet database with information on compound splits. This is restricted to nouns and does not provide filler letters or deep-level structures. The data was revised since then. We are using version 11 which was most recently updated in February 2017.[5] Example (3) presents a typical entry for *Werkstück* 'work piece'. The parts of interest are marked by bold letters. As two derivational processes are possible, two modifiers *werken* 'to work' and *Werk*'work, noun' exist for the head, leading to two splits.

(3) <synset id="s5552" **category="nomen"** class="Artefakt"> <lexUnit id="l8355" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no"> <orthForm>Werkstück </orthForm> <**compound**> <**modifier category="Nomen"**>**Werk**</**modifier**> <**modifier category="Verb"**>**werken**</**modifier**> <**head**>**Stück**</**head**> </**compound**> </lexUnit> </synset>

Different to the CELEX data, the filler letters are missing in the analyses, such as in (4a). Therefore, we insert them by a heuristic method to get analyses as in (4b). Furthermore, we exclude compounds with proper names as constituents such as (5) and foreign expressions as in (6). We did not correct any mistakes of the database but automatically excluded a few deficient entries, for example those with missing part-of-speech classes, and compounds with affixoids or fossilized morphemes.

(4) a. Abfahrtszeit 'departure time': Abfahrt|Zeit (departure|time)
b. Abfahrtszeit 'departure time': Abfahrt|s|Zeit (departure|*filler letter*|time)

(5) Bodenseeregion 'Lake of Constance region'

(6) After-Show-Party

4 Procedures

4.1 Data Extraction

For extracting all relevant information from the refurbished CELEX data, we build an inverted index of all lemmas and extract all immediate constituents and their categories. Then we internally add the infinitive forms of the verbs which are included within these entries. This is necessary so that these forms can be found within the inverted index of the entries. We also 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 like to generate not just compounds and derivatives but also conversions, we extract the relevant information for this word-formation type too but exclude 724 cases of lexicalized inflection (see Gulikers et al., 1995, 54) such as (7).

(7) anhaltend (continuing, present perfect) 'persistent'

[5] see http://www.sfs.uni-tuebingen.de/GermaNet/compounds.shtml#Download for a description.

The data finally comprises an inverted list of 40,081 entries with 38,650 different word splits of complex entries (compounds and derivations) and 1,678 conversions.

From the GermaNet data, we extract all completely annotated compounds with their splits and filler letters according to the restrictions. We also infer the category of the head from the entry. This leads to a list of 64,468 entries with 67,466 different word splits; all of them are nominal compounds. (8) shows the analyses for (3). The variation in spelling of *nomen* is due to the original data.

(8) Werkstück Werk_Nomen|Stück_nomen
Werkstück werken_Verb|Stück_nomen

4.2 Building the Trees

For each entry of the extracted data, the procedure starts from the list of its immediate constituents and recursively collects all information. Algorithm 1 in Appendix A presents the recursive process for the CELEX data, Algorithm 2 for the GN data.

4.3 Diachronic Information

Diachronic information can be of interest, however, for many applications it is considered as unnecessary or even unhelpful. Therefore, the script permits users to choose a threshold of similarity within the range of [0:1] which is compared to a measure we devised based on the Levenshtein distance.

For accepting or rejecting two parts of words, the procedure will calculate the Levensthein distance (LD) for the strings of the smaller length of the two compared constituents $(min(c_1, c_2))$, and then compare their quotient *dis* to a threshold *t* as in (9):

$$dis = \frac{LD}{min(c_1, c_2)} \leq t \tag{9}$$

For calculating the dissimilarity quotient of the example (1), in (10) the stem of the derived form (e.g. *treib*) and its component (e.g. *driften*) are reduced to the smaller size of these forms. In this case, the smaller length is 5. After this, the quotient of *LD* and the length is compared to the threshold. (10) shows that the analysis will stop for a threshold at 0.8 or below.

$$\frac{LD}{min(c_1, c_2)} = \frac{4}{5} \tag{10}$$

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

4.4 Formats of output

The output can be configured in many ways. The following options are available:

- Depth of analysis for compounds
- Parts of speech for the constructs and/or the smallest constituents
- Choice of the output format (parentheses or a notation with | for the splits on the same level)
- Addition of filler letters for GN
- Transfering the GN annotation scheme to CELEX scheme
- Removing compounds with proper names and/or foreign words as constituents for GN
- Analysis of conversions for CELEX
- Depth of analysis for conversions for CELEX
- Dissimilarity measure for CELEX diachronic analyses

The analyses in (11) for (3) and a complex compound containing (3) as a constituent are from the GN part in the format without any linguistic information. Due to combination of ambiguous entries, it comprises multiples trees for some forms. Example (12) shows an output for CELEX data of the same form in parenthesis notation. Here only one analysis is assigned to the word form with the verb as a result of conversion from the noun. More examples are given in the Appendix.

(11)

Werkstück	Werk|Stück	'work(noun)|piece'
Werkstück	werken|Stück	'to work|piece'
Glaswerkstück	Glas|(Werk|Stück)	'glass|work(noun)|piece'
Glaswerkstück	Glas|(werken|Stück)	'glass|to work|piece'

(12) Werkstück (*werken_V* (Werk_N)(en_x))(Stück_N) '(*to work_V* (work_N)(en(suffix))(piece_N)'

5 Merging the trees

The CELEX trees comprise not only compounds but also deep-level analyses of derivatives and conversions, while the GN morphological data is restricted to compound nouns which are partially very complex. For instance, the flat analysis of *Währungsausgleichsfond* 'currency adjustment fond' (13) from the GN database can be recursively augmented to the tree in (14). Its constituent *Ausgleich* 'adjustment' is not further analyzed within the GN database, but has an entry as a complex conversion (15) in the CELEX database. Therefore, the combination of both sets and their parameters for building complex trees seems promising.

(13) Währungsausgleich_N|s_x|Fonds_N 'currency adjustment|*filler letters*|fund'

(14) (*Währungsausgleich_N* Währung_N|s_x|Ausgleich_N)|s_x|Fonds_N)
'(*currency adjustment_N* currency_N|s_x|adjustment_N)|s_x|fund_N'

(15) Ausgleich (*ausgleichen_V* aus_x|(*gleichen_V* gleich_A|en_x))
'adjustment (*to adjust_V aus(*prefix*)_x|(*to equal_V equal_A|en(*suffix*)_x))'

Moreover, GN compounds which were formerly excluded during the procedure of data extraction because their part of speech categories are missing inside the database (see 3.2), can be assigned the category from CELEX if available.

The main problem consists in two annotation sets and their different classification schemes, especially for roots. Table (1) shows the mapping. While the main part-of-speech categories are almost perfectly mappable between the CELEX and the GN data, the classification of function words and bound morphemes is less consistent. There are cases of different interpretations with a tendency of CELEX to prefer affix analyses for cases such as (16) and (17) with *a.* presenting the GN entry and *b.* the entry of CELEX. There are differing analyses of morphological constituency. In (18) GN's compound analysis is opposed to the conversion of CELEX. The classes of roots and word groups have the same or complementary scopes, e.g. (19) and (20) have the same analysis in both sources. We decided to unify the tagset but to leave different trees such as in (11) and (12) to the choice of the users. Some more complex analyses as well as the algorithm are presented in Appendix A.

(16)

a.	Abwasser	(ab_P)(Wasser_N)	'(away_P)(water_N) waste water'
b.		(ab_x)(Wasser_N)	'(away_x)(water_N) waste water'

(17)

a.	afroasiatisch	(afro_R)(Asiatisch_N)	'(afro_R)(Asian_N)'
b.	afroamerikanisch	(afro_x)(amerikanisch_A)	'(afro_x)(American_A)'

(18)

a.	Maßnahme	(Maß_N)(Nahme_N)	'(measure_n)(taking_N) measure'
b.		maßnehmen_V	'(to measure_take_V) measure'

(19) Kondenswasser (kondens_R)(Wasser_N) '(condensed_R)(water_N)'

(20) Zwölftonmusik (zwölf Ton_n)(Musik_N) '(twelve tone_n)(music_N)'

Part of Speech/morph type	GN	CELEX	GermanTreebank
noun	nomen, Nomen	N	N
adjective	Adjektiv	A	A
adverb	Adverb	B	B
preposition	Präposition	P	P
verb	Verb, verben	V	V
article	Artikel	D	D
interjection	Interjektion	I	I
pronoun	Pronomen	O	O
abbreviation	Abkürzung	X	X
word group	Wortgruppe	n	n
root/confix	Konfix	R	R
filler letters, affixes	-	x	x

Table 1: Mapping of two morphological tagsets

6 Results

Table 2 provides the number of the trees for CELEX, GermaNet and their merge in GermanTreebank. The parameters for the deep-level analyses are 6 for the levels of complex words and 2 for conversions. The Levenshtein dissimilarity threshold was set to 0.5. Double entries were removed. As the combinatorial power of GN's ambiguous trees grows with the depth of the trees, the numbers have to be considered with a grain of salt. The set of trees in the GermanTreebank consists of the unification of both sources. For examples, see (21)-(23) in A.

Structures	GN entries	CELEX entries	GermanTreebank
flat	67,452	40,097	100,095
deep-level	68,163	40,097	104,424
merged with CELEX	68,171	n/a	100,986

Table 2: A German Treebank

7 Conclusions and future work

This paper decribes our recent work on merging two types of morphological trees from GermaNet and CELEX. The resulting resource contains 95,506 lemmas connected with 100,986 merged trees and is currently the biggest available data resource of its kind. In principle, the treebank is extensible and combinable with other analyses, and we intend to enlarge it. The resource can be especially useful for all kind of data-intense morphological analyses. We plan to use it especially as a source for depth-level word analyses in combination with a word splitter.

Acknowledgments

The author was supported by the German Research Foundation (DFG) under grant RU 1873/2-1.

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. http://elib.uni-stuttgart.de/opus/volltexte/2014/9768.

Ryan Cotterell, Arun Kumar, and Hinrich Schütze. 2016. Morphological Segmentation Inside-Out Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. The Association for Computational Linguistics, pages 2325–2330. http://aclweb.org/anthology/D/D16/D16-1256.pdf.

Joachim Daiber, Lautaro Quiroz, Roger Wechsler, and Stella Frank. 2015. Splitting compounds by semantic analogy. In *Proceedings of the 1st Deep Machine Translation Workshop*, ÚFAL MFF UK, pages 20–28. http://aclweb.org/anthology/W15-5703.

Matea Filko and Krešimir Šojat. 2017. Expansion of the derivational database for Croatian. In *First Workshop on Resources and Tools for Derivational Morphology (DeriMo)*. http://derimo2017.marginalia.it/index.php/proceedings.

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*, Springer, volume 4002, pages 55–66. https://doi.org/10.1007/11780885_7.

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 *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. pages 9–15. http://www.aclweb.org/anthology/W97-0802.

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*, Niemeyer, Tübingen, pages 53–66.

Nabil Hathout and Fiammetta Namer. 2016. Giving Lexical Resources a Second Life: Démonette, a Multi-sourced Morpho-semantic Network for French. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France. http://www.lrec-conf.org/proceedings/lrec2016/pdf/279_Paper.pdf.

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

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*. Association for Computational Linguistics, pages 187–193. http://www.aclweb.org/anthology/E03-1076.

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*. Association for Computational Linguistics, Berlin, Germany, pages 76–81. http://anthology.aclweb.org/W16-2012.

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)*. Association for Computational Linguistics, pages 617–622. https://doi.org/10.18653/v1/N16-1075.

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). http://www.aclweb.org/anthology/L04-1275.

Elnaz Shafaei, Diego Frassinelli, Gabriella Lapesa, and Sebastian Padó. 2017. DErivCELEX: Development and Evaluation of a German Derivational Morphology Lexicon based on CELEX. In *Proceedings of the DeriMo workshop*. Milan, Italy. http://derimo2017.marginalia.it/index.php/proceedings.

Petra Steiner. 2016. Refurbishing a Morphological Database for German. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, 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). http://www.lrec-conf.org/proceedings/lrec2016/summaries/761.html.

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. http://www.gscl.org/proceedings/2015/GSCL-201508.pdf.

Luigi Talamo, Chiara Celata, and Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure* 9(1):72–102. https://doi.org/10.3366/word.2016.0087.

R. Wahrig-Burfeind and Gütersloh Lexikoninstitut Bertelsmann. 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.

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

Zdeněk Žabokrtský, Magda Sevcikova, Milan Straka, Jonáš Vidra, and Adéla Limburská. 2016. Merging Data Resources for Inflectional and Derivational Morphology in Czech. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France. http://www.lrec-conf.org/proceedings/lrec2016/pdf/994_Paper.pdf.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 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)*. Association for Computational Linguistics, volume 1, pages 1201–1211. http://www.aclweb.org/anthology/P13-1118.

Andrea Zielinski and Christian Simon. 2009. Morphisto –An Open Source Morphological Analyzer for German. In *Proceedings of the 2009 Conference on Finite-State Methods and Natural Language Processing: Post-proceedings of the 7th International Workshop FSMNLP 2008*. IOS Press, Amsterdam, The Netherlands, The Netherlands, pages 224–231. http://dl.acm.org/citation.cfm?id=1564035.1564061.

Patrick Ziering, Stefan Muller, 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*. Association for Computational Linguistics, Berlin, Germany, pages 50–55. http://anthology.aclweb.org/W16-1807.

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*, Association for Computational Linguistics, pages 644–653. http://aclweb.org/anthology/N/N16/N16-1078.pdf.

A Appendix

Formats

The following shows the entries of *Abschlussprüfung* 'final exam', see (2), *Abdrift* 'leeway', see (1), and *Abgangszeugnis* 'leaving certificate', see (3). For all linguistic information, | notation, and a Levenshtein threshold of 0.5, the results are presented in (21), for parenthesis notation and no restrictions on diachronic conversions in (22) and for a flat representation of the immediate constituents see (23).

(21)
```
Abschlussprüfung
(*Abschluss_N*
(*abschließen_V*
ab_x|
schließen_V))|
(*Prüfung_N*
prüfen_V|
ung_x)

Abdrift
ab_x|
(driften_V)

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

(22)
```
Abschlussprüfung
(*Abschluss_N*
(*abschließen_V*
(ab_x)
(schließen_V)))
(*Prüfung_N*
(prüfen_V)
(ung_x))

Abdrift
(ab_x)
(*driften_V*
treiben_V)

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

(23)
```
Abschlussprüfung
Abschluss_N|
Prüfung_N

Abdrift
ab_x|
driften_V

Abgangszeugnis
Abgang_N|
s_x|
Zeugnis_N
```

Example of the Treebank: Augmentation of a GermaNet tree

The following shows how an entry from GermaNet (GN), *Währungsausgleichsfonds* 'currency adjustment fund' (24) can be augmented recursively from GN (25) and the CELEX database (26). The complete tree is presented in (4).

(24) Währungsausgleich_N|s_x|Fonds_N 'currency adjustment|*filler letters*|fund'

(25) (*Währungsausgleich_N* Währung_N|s_x|Ausgleich_N)|s_x|Fonds_N
'(*currency adjustment_N* currency_N|s_x|adjustment_N)|s_x|fund_N'

(26) (*Währungsausgleich_N* Währung_N|s_x|(*Ausgleich_N*
(*ausgleichen_V* aus_x|(*gleichen_V* gleich_A|en_x))))|s_x|Fonds_N
'(*currency adjustment_N* currency_N|s_x|(*adjustment_N*
(*to adjust_V* aus,Prefix_x|(*to equal_V* equal_A|en_x))))|s_x|fund_N'

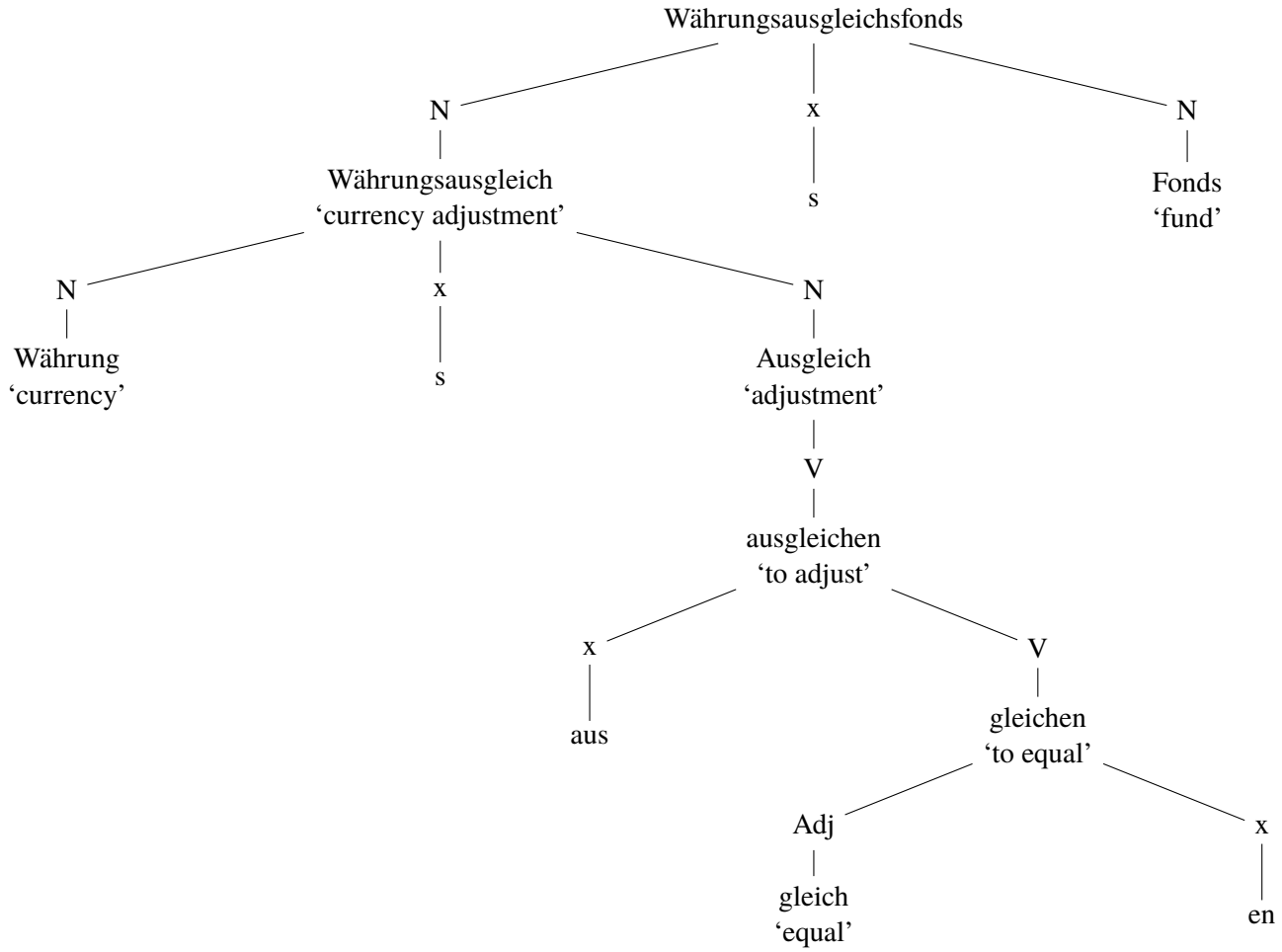

Figure 4: Merged morphological analysis of *Währungsausgleichsfonds* 'currency adjustment fund'

Algorithms

Algorithm 1: Building a morphological treebank from CELEX German data

```
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++;
                    analysedeepercelex part with parameters and depth;
                    return result of analysedeepercelex
                end
            end
        end
    end
end

 sub analysedeepercelex 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
        analysedeepercelex subpart
          if levenshtein threshold and analysedeepercelex subpart is dissimilar then
            skip deeper analysis;
            return subpart
        end
        else
            return result of analysedeepercelex subpart
        end
    end
end
```

Algorithm 2: Building a morphological treebank from GermaNet flat compounds

```
Input: GN flat compounds
Output: A Morphological Treebank
initialization of parameters: depth of analysis, linguistic information, parts of speech, style of
  output;
forall entries of GN flat compounds do
    if entry is a compound
     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
    depth of analysis++;
    foreach subpart of part do
        analysedeeper subpart
          return result of analysedeeper subpart
    end
end
```

Algorithm 3: Building a merged morphological treebank from GermaNet and CELEX

```
Input: CELEX-German revised, GN flat compounds
Output: A Morphological Treebank
initialization of parameters: depth of analysis, linguistic information, levenshtein threshold, parts
  of speech, style of output;
add CELEX data to the knowledge base
 forall entries of GN flat compounds do
    if entry is a compound then
        foreach constituent of entry do
            if depth of analysis reached then
                retrieve linguistic information/PoS as required;
                return linguistic information and constituent
            end
            else if constituent not found in GN data then
                depth of analysis++;
                analysedeepercelex as in Algorithm 1 part with parameters and depth;
                return result of analysedeepercelex
            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 if constituent not found in GN data then
    depth of analysis++;
    analysedeepercelex as in Algorithm 1 part with parameters and depth;
    return result of analysedeepercelex
end
else
    depth of analysis++;
    foreach subpart of part do
        analysedeeper subpart
          return result of analysedeeper subpart
    end
end
```