Converting a Database of Complex German Word Formation for Linked Data

Petra Steiner
Universität Bayreuth
petra.steiner@uni-bayreuth.de

Abstract
This work combines two lexical resources with morphological information on German word formation, CELEX for German and the latest release of GermaNet, for extracting and building complex word structures. This yields a database of over 100,000 German wordtrees. A definition for sequential morphological analyses leads to an Ontolex-Lemon type model. By using GermaNet sense information, the data can be linked to other semantic resources. An alignment to the CIDOC Conceptual Reference Model (CIDOC-CRM) is also provided. The scripts for the data generation are publicly available on GitHub.

Keywords: CELEX, GermaNet, morphology, German

1. Introduction
Languages with a high lexical productivity in word formation bounce into bottleneck problems if it comes to analysing texts, building terminologies, or finding links between ontologies and other networks. Concerning the German language, there are three main problems:

A. The wealth of ambiguous forms on the level of word formation
B. The lack of deeper structural analyses in current approaches
C. The lack of linkages between morphological analyses and ontologies

The linkage of lemmas, lexical items, ontological entities etc. with morphological complex word forms presupposes their structural disambiguation on the morphological level, either manually or automatically. Only if this is provided, a classification at a high quality level is possible. However, especially for long and complex lexical items, the morphological analyses and with it the semantic interpretations are no trivial task for human and automatic disambiguation.

For example, Landesentwicklungsgesellschaft ‘state development corporation’ and Stadtentwicklungs- gesellschaft ‘urban development company’ have two different hyperonyms although their first constituents Land ‘state’ and Stadt ‘urban’ are cohyponyms denoting levels of administrative units. However, the first term denotes a corporation, and the second a company, as the German lexeme Gesellschaft can be used for both senses. Figure 1 and Figure 2 present the first three levels of the different structures, including the linking elements. The last top level constituents of the morphological structure (here Entwicklungsgesellschaft vs. Gesellschaft) are usually the heads of the constructions, especially for compounds. By this, they determine not only the grammatical features of the complete lexeme but in most cases also the hyperonymic class of the terms.

German compounds can consist of derivatives such as Entwicklung and Gesellschaft, both ending with suffixes (ung and schaft). These analyses can further link lexical units to others, e.g. by the verbs they were derived from. On each level of morphological segmentation, the number of possible analyses is $2^n$. This number can be reduced by excluding implausible constructions such as suffixes at the beginning of a construct. However, it has to be multiplied by the number of morphological homonyms for the segmented forms. The wealth of such long and structurally ambiguous wordforms necessitates the search for solutions.

This paper provides the development of a lexical resource for complex morphological analyses. Section 2 gives a concise overview of related work in word seg-
3. Lexical Resources for the Synopsis of Morphological Analyses

3.1. The Refurbished CELEX-German Database

CELEX is a publicly available database of Dutch, English, and German lexical information (Baayen et al., 1995). The German part of the CELEX database (CELEX-German) comprises 51,728 lemmas of all parts of speech. 38,650 entries are derivates or compounds and 2,402 entries are conversions. The compilation of the lemmas is widely overlapping with the one of the dictionary Der kleine Wahrig (Wahrig-Burfeind and Bertelsmann, 2007) which represents the core vocabulary for German. CELEX-German comprises not just flat analyses but also German word tree information. The linguistic information is combined with frequency information based on corpora (Burnage, 1995) which makes it useful for automated morphological and phonological analysis of unknown words. Therefore, CELEX-German (Baayen et al., 1995) is a solid standard for building morphological resources.

The drawbacks of the German part of the CELEX database are its outdated format and the use of former orthographical conventions. Therefore, both lemmas and word forms are transferred to a modern standard of encoding by merging the orthographic and the morphophonological information, both for the lemma and the word form data (Steiner, 2016). After these changes, the database with its solid list of base vocabulary yields a foundation for further exploitation. It serves as the foundation for the morphological structure database and can then be augmented by other resources (Steiner and Ruppenhofer, 2018; Steiner, 2017; Steiner, 2019a; Steiner, 2019b), the first of which is the GermaNet database which contains markup for compounds.

Some of the morphological analyses of the CELEX-German database on a deep level are oriented towards diachronic descriptions. For instance, *Gift* ‘poison’ is analyzed as a derivation from *geben* ‘give’. This is certainly of less interest for linking semantic information. On the other hand, the relation between *Ausfuhr* ‘export.n’ and *ausführen* ‘to export’ is morphologically manifested in an implicit derivation with a/u ablaut and might lead to interesting connections.

The refurbished database possesses no modification concerning this feature. The decision whether to appreciate, accept, or change this diachronic information is left to the next steps of usage, depending on the respective application.

Examples 1 and 2 show parts of the entries for the derivatives *Entwicklung* ‘development’ and *Gesellschaft* ‘society, corporation, company’ with the affixes ent, ung, and schaff

1. *Entwicklung* entwickel+ung[Vx][V...]
   (((ent)[V],V),((Wickel)[N][V])[V],
   (ung)[N][V,][N]]

2. *Gesellschaft* geschaff+Schaft

In most cases, also German morphological data resources are restricted to lists of flat analyses, for instance, the test set of the 2009 workshop on statistical machine translation, which was used by (Cap (2014). 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. Shafaei et al. (2017) use the CELEX 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 (Baayen et al., 1995).
3.2. Compound Analyses from GermaNet

Henrich and Hinrichs (2011) augmented the GermaNet database with information on compound splits. This feature is restricted to nouns. We are using version 17 which was most recently updated in April 2022. This version includes 205,000 lexical units. GermaNet comes with an alignment to Wiktionary entries (Henrich et al., 2011) and connects its senses to EuroWordNet by an interlingual index.

Example 3 and 4 present the entries for Landesentwicklungsgesellschaft ‘state development corporation’ and Stadtentwicklungsgesellschaft ‘urban development company’. The first entry has the hyperonym {Amt, Behörde} ‘office, authority’. The parts of interest are marked by bold letters.

Example 2 presents the entries for Gesellschaft gesell+schaft(Vx)[...]

As can be seen, these entries do neither provide filler letters, such as es or S, nor deep-level structures. Again, it is left the next steps of usage to appreciate, accept, or change this information.

4. Procedures

In general, the underlying script permits to restrict the analysis to GermaNet data. Here, both databases are to be combined.

4.1. Fitting the CELEX Data

For the peculiarity of the CELEX database with its di-achronically motivated derivations, we added a heuristic based on the Levenshtein distance. For accepting or rejecting two parts of words as derivational relatives, the procedure will calculate the Levenshtein distance (LD) for the (sub)strings of the smaller length of the two compared constituents (min(c1, c2)), and then compare their quotient dis to a threshold t as in (5):

\[
\text{dis} = \frac{\text{LD}}{\text{min}(c_1, c_2)} \leq t
\]  

(5)

For example, for the derivational pair Gift - geb, the smaller length is 3. The string Gift is cut to this length: Gif. After this, the quotient of LD for Gif and geb and the length is compared to the threshold. (6) shows that the analysis will stop for a threshold at 0.66 or below.

\[
\frac{\text{LD}}{\text{min}(c_1, c_2)} = \frac{2}{3}
\]  

(6)

4.2. Fitting the GermaNet Data

Different to the CELEX data, the filler letters in the GermaNet data are missing within the analyses. A heuristic method recovers them. A few entries were automatically excluded, as those with missing part-of-speech classes which could not be retrieved from the CELEX database, and compounds with affixoids or fossilized morphemes. Complex components whose analyses are not inside the database are considered as technically simplex lexemes.

4.3. Synopsis of the Databases

The structures are recursively collected, first from the GermaNet data and if no entries can further be found there, then CELEX-German with its rich information on derivations is retrieved. By this, compositional constituents not found within the GermaNet inventory but inside CELEX-German can be analyzed too. Algorithm I presents the top-down procedure. Among others, the underlying program has the options presented in Table 1.

We permit compounds with proper names as constituents and foreign expressions, automatically add filler letters and choose a threshold of 0.5 for dissimilarity. Parts of speech tags of GermaNet and CELEX-German are mapped according to Table 2. In GermaNet, there are some orthographic variants of these categories, e.g. nomen and Nomen for noun. The chosen depth for constructions of conversions is 2 and the general depth for the trees is 7, as a depth of 8 did not yield any deeper analyses.

The GermaNet Release 17.0 yields 97,362 compounds, including some with proper names and foreign words as
**Input:** CELEX-German revised, GN flat compounds

**Output:** A Morphological Treebank initialization of parameters: depth of analysis, linguistic information, levenshtein threshold, parts of speech, filler letters, conversions (Zusammenrückungen), style of output;

**add CELEX data to the knowledge base according to the requirements**

forall entries of GN flat compounds do
  if entry is a compound according to the conditions (complete parts of speech, foreign words, proper names yes/no) 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 part with parameters and depth;
        return result of analysedeepercelex
      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 if constituent not found in GN data then
    depth of analysis++;
    analysedeepercelex part with parameters and depth;
    return result of analysedeepercelex
  end
  else
    depth of analysis++;
    foreach subpart of part do
      analysedeepercelex subpart
      return result of analysedeepercelex subpart
    end
  end
end

**Algorithm 1:** Building a merged morphological treebank from GermaNet and CELEX

- **-rmfw** ignore lexemes with foreign expressions
- **-rpmn** ignore lexemes with proper names
- **-addfl** add filler letters
- **-n** iterations for the depth of tree for compounds and derivations
- **-zn** iterations for the depth of conversions in CELEX
- **-levperc** Levenshtein based threshold, range 0:1
- **-celex** use CELEX compounds and derivations
- **-zcelex** use CELEX conversions
- **-ctags** map GermaNet tags to CELEX tags
- **-pos** provide parts of speech
- **-par** choose parenthesis style for the output

**Table 1: Options for Linking the Databases**

components but excluding all lexemes with affixoids or fossilized morphemes. The number of deep-level analyses amounts to 119,476.

As examples, the complete analyses of our examples are presented in 7 and 8. Table 3 shows the number of entries for the merged databases, some of them are alternatives for ambiguous parts.

(7) **Landesentwicklungsgesellschaft**
(Land_N)
(es_x)
(*Entwicklungsgesellschaft_N*
(*Entwicklung_N*
(*entwickeln_V*
(ent_x)
(*wickeln_V*
(Wickel_N)(n_x)))))
(ung_x))
(s_x)
(*Gesellschaft_N*
(gesellen_V)
(schaft_x)))

(8) **Stadtentwicklungsgesellschaft**
(*Stadtentwicklung_N*
(Stadt_N)
(*Entwicklung_N*
(*entwickeln_V*
(ent_x)
(*wickeln_V*
(Wickel_N)
(n_x))))
(ung_x))
(s_x)
(*Gesellschaft_N*
(gesellen_V)
(schaft_x)))
5. Linkages

5.1. Linking Morphological Data to Ontolex-Lemon

Ontolex-Lemon (McCrae et al., 2017) can be considered as the main standard for lexical data on the web. Its core component was tailored for linking ontologies with resources of lexical entries, consisting of information of sense and form. Declerck and Racioppa (2019) and Racioppa and Declerck (2019) provide information concerning inflection of word forms. However, standards for the description of (complex) morphological analyses are still under development (Klimek et al., 2019). Morph classes such as affix or prefix are insufficient for describing structures which are not just defined by hierarchy but also by sequence. Therefore, representing constituency by decomp:Component and decomp:Constituent (Klimek et al., 2019, 585ff.) resources could be accompanied by next markers for making the level and the position of the relation transparent. A next element is easily definable by rdf:rest (the next element is the first element of the rest). Another option is using expressions of one-level sets with fixed sequence. rdf:seq (https://www.w3.org/TR/rdf-schema/#ch_seq) provides this feature, as it is an ordered container. Listing 9 displays the lemma Landesentwicklungsge- sellschaft with such an analysis.

```
(9) lexinfo:orderedAnalysis a rdf:seq;
    rdfs:comment "A list of ordered components as defined by decomp:Component";
    rdfs:range :decomp:Component;
    rdfs:subPropertyOf lexinfo:morphosyntacticProperty.
```

5.2. Linking Morphological Data to CIDOC

The derived morphological information is intended to be used to link information of cultural heritage. Therefore, it can aligned to the CIDOC Conceptual Reference Model (CIDOC-CRM) (Mambrini and Passarotti 2020) establish the linkage to CIDOC-CRM via the propositional status of etymological assumptions. In case of morphological analyses, the class E33 Linguistic_Object (for the definition, consult https://cidoc-crm.org/Entity/e33-linguistic-object/version-6.0) is more suitable in analogy to Wettlauer et al. (2015, 191ff.).
Figure 3: An model for the representation of morphological and semantic information of *Landesentwicklungsge-
ellschaft* ‘state development corporation’
5.3. Senses and Synopsis
As minimal linguistic signs, morphemes have meanings and/or functions. As GermaNet provides the synsets for the components of the morphological analyses, the connection to their content side is straightforward. The inventories of the Interlingual Index to EuroWordNet and of the aligned Wiktoryan resources open the way to Linked Open Data (Chiarosc et al., 2020). Figure [illustrates a synopsis of these connections. For the sake of clarity, some relations were omitted.

6. Conclusions and Future Work
This paper links the most recent version of GermaNet with the established resource of CELEX-German by recursively connecting their compositions, conversions and derivations, and mapping the annotation sets. Furthermore, it takes a step towards the representation of sequential and hierarchical morphological information for Ontolex-Lemon and similar models by using the rdfs:Container class Seq which is defined as an ordered list.

Finally, a transparent connection to CIDOC-CRM is provided to make this linguistic data findable, accessible, interoperable, and reusable for other applications, in the sense of the FAIR data principles (Wilkinson et al., 2016).

The information of the linguistic databases can be considered as on a high-quality level. However, as the inventories of both lexical resources are restricted, hybrid approaches with (more time-consuming) morphological parses and enrichments of the knowledge base are one of the next steps (Steiner, 2019a) for the linguistic work. This would also help to find candidates within the database which could get a more fine-grained analysis. Especially, for new entries whose components are not yet parts of the data, this can be useful. Another very important step will connect the morphological analyses to ontological knowledge via the WordNet synsets by direct mappings of the interlingual and Wiktionary entries.

The scripts for the data generation are publicly available on https://github.com/petrasteiner/morphology

7. Acknowledgements
Work for this paper was partially supported within the Africa Multiple Cluster of Excellence at the University of Bayreuth, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC2052/1 – 390713894.

8. Bibliographical 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.

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.

Chiarosc, C., Klimek, B., Fäth, C., Declerck, T., and McCrae, J. P. (2020). On the linguistic linked open data infrastructure. In Proceedings of the 1st International Workshop on Language Technology Platforms, pages 8–15, Marseille, France. European Language Resources Association.

Declerck, T. and Racioppa, S. (2019). Porting multilingual morphological resources to ontolex-lemon. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pages 233–238, Varna, Bulgaria. INCOMA Ltd.

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). GERT-WOL und morphologische Disambiguiierung 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 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 Morpholimics 1994, pages 53–66. Niemeyer, Tübingen.

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.

Henrich, V., Hinrichs, E. W., and Vodolazova, T. (2011). Aligning GermaNet senses with Wiktoryan sense definitions. In Zygmunt Vetulani et al., editors, Human Language Technology Challenges for Computer Science and Linguistics - 5th Language and Technology Conference, LTC 2011, Poznań, Poland, November 25-27, 2011. Revised Selected Papers, volume 8387 of Lecture Notes in Computer Science, pages 329–342. Springer.

Klimek, B., McCrae, J. P., Bosque-Gil, J., Ionov, M., Tauber, J. K., and Chiarosc, C. (2019). Challenges
for the representation of morphology in ontology lexicons. Proceedings of eLex, pages 570–591.

Mambri, F. and Passarotti, M. (2020). Representing etymology in the LiLa knowledge base of linguistic resources for latin. In Ilan Kernerman, et al., editors, Proceedings of the 2020 Globlex Workshop on Linked Lexicography, pages 20–28, Marseille, France. European Language Resources Association.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitemaar, P., and Cimiano, P. (2017). The OntoLex-Lemon model: Development and applications. In Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. Leiden, the Netherlands, 19–21 September 2017, pages 587–597.

Racioppa, S. and Declerck, T. (2019). Porting the latin wordnet onto ontolex-lemon. In Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography, pages 429–439.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In Proceedings of the International Conference on Language Resources and Evaluation (LREC).

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. (2018). Building a morphological treebank for German from a linguistic database. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May. European Language Resources Association (ELRA).

Steiner, P. (2016). Refurbishing a morphological database for German. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 1103–1108, Portorož, Slovenia, May. 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, pages 146–160, Prague, Czech Republic.

Steiner, P. (2019a). Augmenting a German morphological database by data-intense methods. In Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 178–188, Florence, Italy, August. Association for Computational Linguistics.

Steiner, P. (2019b). Combining data-intense and compute-intense methods for fine-grained morphological analyses. In Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology, pages 45–54, Prague, Czechia, 19–20 September. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

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 grammatikalen und orthografischen Aspekten der deutschen Gegenwartssprache]. Wissen Media Verlag.

Wettlaufer, J., Johnson, C., Scholz, M., Fichtner, M., and Thotempudi, S. G. (2015). Semantic Blumenbach: Exploration of text–object relationships with semantic web technology in the history of science. Digital Scholarship in the Humanities, 30(Supplement 1):fqv047.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., and Appleton, G. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific data, 3:160018.

Würzner, K. and Hanneforth, T. (2013). Parsing morphologically complex words. In Mark-Jan Nederhof, editor, 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. The Association for Computer Linguistics.

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.