Published in: Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry; Goggi, Sara; Grobelnik, Marko; Maegaard, Bente et al. (Hg.) (2016): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), pp. 1103-1108.

Refurbishing a Morphological Database for German

Petra Steiner
Institute for Information Science and Natural Language Processing
Hildesheim University
steinerp@uni-hildesheim.de

Abstract

The CELEX database is one of the standard lexical resources for German. It yields a wealth of data especially for phonological and morphological applications. The morphological part comprises deep-structure morphological analyses of German. However, as it was developed in the Nineties, both encoding and spelling are outdated. About one fifth of over 50,000 datasets contain umlauts and signs such as *ß*. Changes to a modern version cannot be obtained by simple substitution. In this paper, we shortly describe the original content and form of the orthographic and morphological database for German in CELEX. Then we present our work on modernizing the linguistic data. Lemmas and morphological analyses are transferred to a modern standard of encoding by first merging orthographic and morphological information of the lemmas and their entries and then performing a second substitution for the morphs within their morphological analyses. Changes to modern German spelling are performed by substitution rules according to orthographical standards. We show an example of the use of the data for the disambiguation of morphological structures. The discussion describes prospects of future work on this or similar lexicons. The Perl script is publicly available on our website.

Keywords: morphology, word structure, deep-structure morphological analyses, CELEX, orthography

1. Introduction and related work

While the number of existing digital linguistic data increases, sustainability is growing more and more important within the field of language resources. Whereas some approaches favor the exploitation of multi-authored and distributed knowledge such as Wiktionary (Sagot, 2014; Sylak-Glassman et al., 2015), others' work also uses older and more established sources (Borin et al., 2009). The current project aims at updating a part of CELEX, which is a database of Dutch, English, and German lexical information (Baayen et al., 1995). Besides information on orthographic, phonological and syntactic features, it also contains ample information on word-formation, especially manually annotated multi-level word structures.

Current morphological analyzers for German yield flat structures of morphs or word constituents on different levels, e.g. SMOR (Schmid et al., 2004), Gertwol (Haapalainen and Majorin, 1995), MORPH (Hanrieder, 1991; Hanrieder, 1996), TAGH (Geyken and Hanneforth, 2006) but no hierarchical parses which can provide important information for word sense disambiguation. Only Würzner and Hanneforth (2013) present an approach for full morphological parsing of German, which is, however, restricted to adjectives.

By contrast, the German part of the CELEX database comprises word tree information for a lexicon containing words of all parts of speech and is therefore an important source for deep-structure 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.

However, the German part of the database has some drawbacks which impair its usefulness: About one fifth of the 51,728 lemmas contain letters such as *ä* or *ß*. As the database was created at a time before modern encoding, umlauts are represented as *ae*, *oe* etc. Unfortunately, these strings cannot be recovered by simple substitutions, as (1) demonstrates, while for (2) the replacement of an umlaut is required.

(1) Lemma: *Oboe* - representation *Oboe* 'oboe'

(2) Lemma: *böse* - representation *boese* 'bad'

Also, while the orthographic information for the lemmas of headwords and stems can be extracted easily from the lexicon, this is not the case for some parts within the morphological analyses.

Another problem is the use of an out-dated spelling convention which makes the lexicon partially incompatible with text written after 1996 when spelling reforms were implemented in Austria, Germany and Switzerland. For instance, the modern spelling of the originally CELEX entry *Abschluß* 'conclusion' is *Abschluss*.

As the database was created according to the standardized spelling conventions of its time, there are only a few spelling mistakes which call for corrections and no spelling variants as in older historical German texts. The changes are very regular and do not have to be tackled by using variants of Levenshtein Distance as done by Bollmann et al. (2012).

Section 2 describes the German part of the CELEX database with an emphasis on the data which are relevant for the updating process. Section 3 presents the procedure we used for the changes. It starts with the change to a modern encoding by merging orthographic and morphological information from the database, followed by revisions of the morphological analyses. The last part consists in transferring the database from old to new German spelling. The results of the script are presented in Section 4. Section 5 shows an example of the use of the data for disambiguating morphological structures. Problems and future directions are discussed in Section 6.

2. German in the CELEX Database

For each of the languages of the CELEX Database (Baayen et al., 1995), three types of linguistic information are provided: orthographic data, information on word formation, and syntactic information. For our purposes, only the first two parts for the German language are of interest.[1] The orthographic information (German orthography lemmas, henceforth GOL) contains information on umlauts and other special characters, corpus frequency and syllabic segmentation for headwords and stems. Special characters are represented by diacritical marks or other symbols. (3) shows an extract of a typical entry with such an encoding. Morphological information (German morphology lemmas, henceforth GML) as in (4) comprises, among others, the corpus frequency, the word-formation type, the singular and plural inflectional patterns, and the immediate constituents. For *Abschlußprüfung* 'final exam' the immediate constituents are *Abschluß* 'conclusion' and *Prüfung* 'exam'. Furthermore, complete morphological parses are provided, as in (4), which is rendered in tree form in Figure (1).

(3) 605\Abschlu$pr"ufung\14\Ab" schlu$" pr"u" fung \N\Abschlu$pr"ufung\Ab-schlu$-pr"u-fung\N

(4) 605\Abschlusspruefung\Abschluss+Pruefung\ ((((ab)[V|.V],(schliess)[V])[V])[N], ((pruef)[V],(ung)[N|V.])[N])

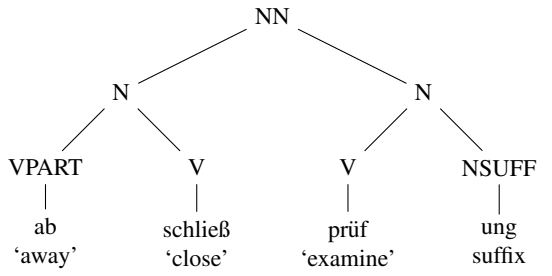

Figure 1: Morphological analysis of *Abschlußprüfung* 'final exam'

In GML, non-ascii letters are represented by ascii substitutes such as *ae* for *ä* or *ss* for *ß*. While the information about these special characters of a lemma can be easily extracted from the related entry in GOL, this is not always possible for the morphological analysis of the deepest morphological layer. In (4), *Abschluß* is a derivative of the verb *schließen* 'to close'. It is a result of ablaut alternation and by this leads to spelling variants of *ss* vs. *ß*, at least in Germany and Austria since the orthographic reforms.[2]

3. Revision of the data

In most cases, the orthographic information of lemmas and their constituents could be retrieved from GOL, as the headwords of verb stems such as *schließ* can be retrieved. However, for root formations such as *fötal* 'fetal' with the segmentation *(föt|al)* there is no entry for the root form *föt* (here spelled as *foet*). A simple substitution to umlauts is not possible, as other roots such as *aero* 'aero' do not require this change.

Therefore, in order to avoid time-consuming post-editing, the components of morphological analyses have to be adapted by heuristics and manually checked. Finally, the lexical database can be generated according to the most recent spelling rules. Figure 2 provides an overview of the procedure for changing the German morphology part of the CELEX database to modern standards. The implementation was done in Perl 5.14 on Linux.

3.1. Merging orthographic and morphological information

The procedure starts with processing pair-wise entries from the orthographic data (GOL) and the morphological part of the database (GML). If the orthographic representation contains one or more diacritics, the generation of the modern encoding is called. If there is just one diacritic character within the word, the character substitution is trivial and also performed on its morphological analyses adjusting for variation between upper-case and lower-case initials of constituents and headwords.

Otherwise, the surrounding characters are added to the replacement patterns. This prevents incorrect substitutions for words such as *Zuschauertribüne* 'grand stands' with two strings of *ue* in the morphological database entry *Zuschauertribuene*, of which only one has to be changed. If however the contexts of one character are equal too, as for ***Ausss**chuss* 'commission', where only the second *ss* is to be transformed according to the old spelling rules, the lemmas are added to a control output file, to be manually checked and, if necessary, treated in the next step. For the current state of the database this holds for only 14 entries. All of them can be transformed in the correct way for a context of two characters preceding (for *ss*) or following the ambiguous string.

3.2. Changing morphs within morphological analyses

The produced datasets with the revised lemmas and some of the morphological analyses are used as input for the changes of the morphological analyses. Some morphs in the analyses contain diacritics while the lemmas do not. For example, in (5) the first immediate constituent of *Singularitaet (Singularität)* 'singularity' is the adjective *singulaer (singulär)* 'singular/unique'.

(5) Singularitaet ((singulaer)[A],(itaet)[N|A.])

For other morphs, lists of substitution rules have to be compiled. For instance, while the character sequence *oes* cannot be generally converted to *ös*, that change is always appropriate when it concerns the suffix *ös*.

(6) Generositaet (((gener)[R],(oes)[A|R.])[A], (itaet)[N|A.])

[1] For an exhaustive description of the database see Gulikers et al. (1995).

[2] The spelling conventions for Swiss German prescribe that *ss* is used instead of *ß*.

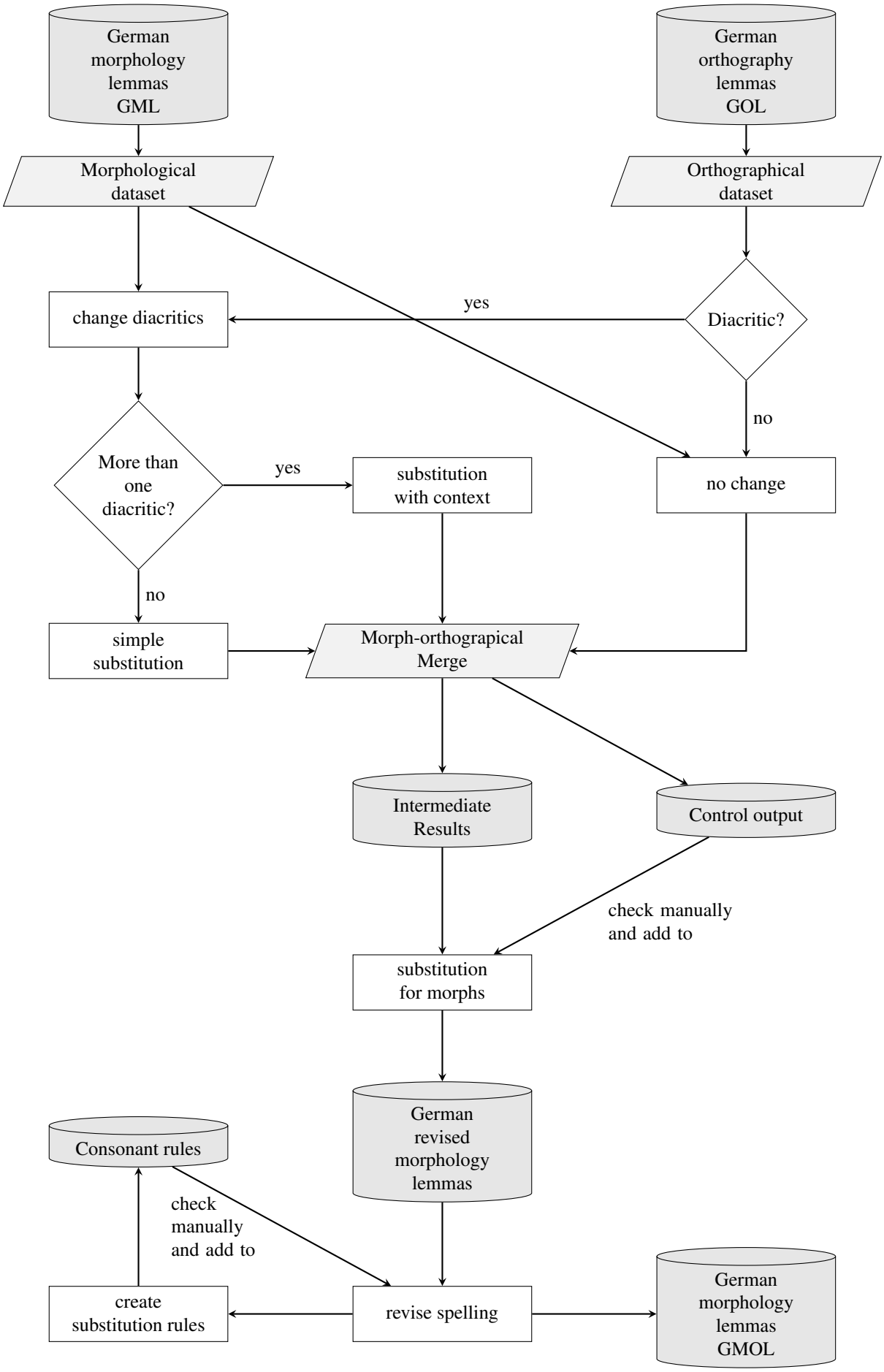

Figure 2: Transfer of the German morphology data to modern standards

Moreover, certain typographical errors can be corrected in this step as well as some overgeneralizations from the first step. The procedure contains 257 substitution rules of words and strings.

3.3. Changes to modern German spelling

The old spelling can be changed to the modern spelling version of German as it is used in Germany. This in particular concerns words with short vowels preceding former instances of *ß* as in (7). Another difference results from the former rule to delete one of three identical consonants if no different consonant follows as in (8). The words concerned are usually compounds.

(7) a. *Prozeß* 'prozess' (old spelling)
b. *Prozess* (new spelling)

(8) a. *Schiffahrt* 'shipping (ship ride)' (old spelling)
b. *Schifffahrt* (new spelling)

88 substitution rules for the first type of change were derived according to the rules of Dudenredaktion (2013).
The lemmas for which the second rule applies are difficult to identify in a list of over 50,000 entries. Therefore, the substitution rules for double to triple consonant clusters are derived semi-automatically and incrementally from the morphological part of the lexicon. For example, in (9) the numbers of *f* in the lemma and immediate constituents differ.

(9) Schiffahrt - Schiff+Fahrt

This leads to the production of the substitution rule in (10)

(10) $stransformed =~ s/Schiffahrt/Schifffahrt/g;

After the new spelling variant *Schifffahrt* has been entered into the substitution rules, the old variant *Flussschiffahrt* 'river navigation' can be found in the next step as a candidate for another substitution rule, as the numbers of *f* in the lemma and immediate constituents differ now (11). This entry is transformed in the next step.

(11) Flussschiffahrt - Fluss+Schifffahrt

This procedure is manually supervised and continues until no further substitution rules can be found. We refer to the refurbished data as GMOL.

4. Results

10,106 of the 51,728 GML entries contain diacritics, in most cases within both the lemma and their morphological analyses. The merging of GOL and GML information leads to the revision of 9,980 entries with 9,072 umlauts (e.g. *ä*), 1,491 instance of *ß* and 9 letters with acute accent. The procedure for revising the morphological analyses yields 613 changes. All of them are correct. The adaption of the modern spelling to the headwords and the components of their morphological analyses leads to 576 updated entries, of which 526 involve changes of *ß* to *ss*.

Three cycles of the generation of double consonants rules yield 41 substitution rules and 46 changes within the entries. All in all, 10,197 entries from the morphological database were updated.
We tested GMOL on the 1,101 lexical items of Cap's (2014, 95) gold standard for the task of compound splitting, who uses part of the test set of the 2009 Workshop on statistical machine translation.[3]
Of these, we created a list of (best) morphological analyses produced by a combination of SMOR and an heuristical approach (Steiner and Ruppenhofer, 2015). We extracted all types of constituents and compared this list with all constituents of GML's respectively GMOL's morphological analyses. Table 1 provides an overview of our investigation.

	Sum	GML	GMOL
Types	1265	1010	1111
Overall Recall		0.80	0.88
Unfound diacritics	121	121	11
Recall for diacritics		0	0.91
Simple substitution	121	36	36
Recall for simple subst.		0.70	0.70

Table 1: Recall for the old and the refurbished database

The eleven constituents with the unfound diacritics were either missing in the CELEX database (e.g. *Ära* 'era') or wrongly analysed parts of syntagmatic compounds which are erroneously segmented as endocentric compounds such as *jährig* '*year-ig(suffix)' of *50-jährig* '50-years old'. If a simple substitution of all instances of *ae* to *ä* etc. is used, the wrongly generated forms would produce a recall of 0.70 for the forms including diacritics and 0.86 for all items. As 256 words of the gold standard comprise diacritics, the impact on the word level is actually higher.

5. Application

The morphological database can be used to obtain reliable frequency information on morphs and constituents. Table 2 presents some of the most frequent morphs in the database.

f(m)	m	f(m)	m	f(m)	m
3588	ung	...	...	896	ab
3066	er	983	aus	845	an
2327	s	974	keit	831	isch

Table 2: Frequencies of morphs *m* from GMOL

This kind of information can be used for shallow and deep-level analyses of morphological structures as done in (Steiner and Ruppenhofer, 2015). Here, we took the geometric mean shown in (12) as a quality measure for selecting one out of a candidate set of multiple possible segmentations.

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} \quad for \quad x_i ... x_n, \tag{12}$$

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

As frequencies we used the counts of morphs and immediate constituents of the revised German CELEX data. Figures (3) and (4) give two possible analyses of the German compound *Anbaumenge* 'cultivation amount'. Only the first tree represents a sensible decomposition.

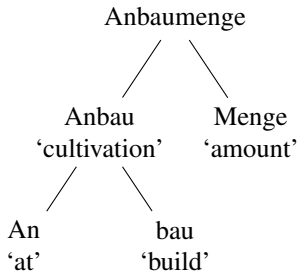

Figure 3: Correct analysis of *Anbaumenge*

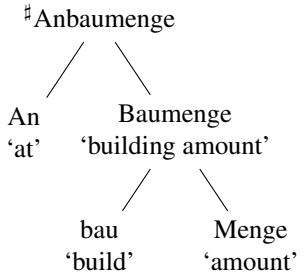

Figure 4: False analysis of *Anbaumenge*

The numbers of the constituents of *Anbaumenge* as retrieved from the CELEX database are $x_1 = 366$ for *an*, $x_2 = 53$ for *bau*, $x_3 = 8$ for *Menge*, $x_4 = 7$ for *Anbau*, and $x_5 = 0.1$ for *Baumenge*, as this is the heuristical value of strings for which no entry in the CELEX database could be found. The values of the geometric means are $gm(An|bau|Menge) = 53.74$ and $gm(An|Baumenge) = 6.05$. This leads to the preference for the first analysis. Steiner and Ruppenhofer (2015, 56) find an overall recall of 81.11 and a weak recall of 96.15 for such weighted decisions.

6. Discussion and future prospects

Changing one fifth of a German morphological database according to modern encoding and spelling yields a wealth of 38,397 morphological analyses on the level of immediate constituents as well as deeper analyses, and information on exactly 200 different word-formation types. As the CELEX Database (Baayen et al., 1995) is under license from the European Language Resources Association, the revised version cannot be made publicly available. However, the script for the refurbishment of the data can be downloaded at `https://www.uni-hildesheim.de/media/fb3/informationswissenschaft/IWiSt-CL/Steiner/OrthCELEX.pl`.

Some drawbacks of the data are:

- The *Mannheim Corpus* which was used for the frequency counts (Gulikers et al., 1995, 102ff.) is rather small.
- Some of the derivations have a strong emphasis on diachronic derivation as shown for *Verschlüsselung* 'encryption' in Figure (5) where the noun *Schlüssel* 'key' is analysed as a derivation from the verb *schließen* 'close' which is certainly justified but not adequate for every task.

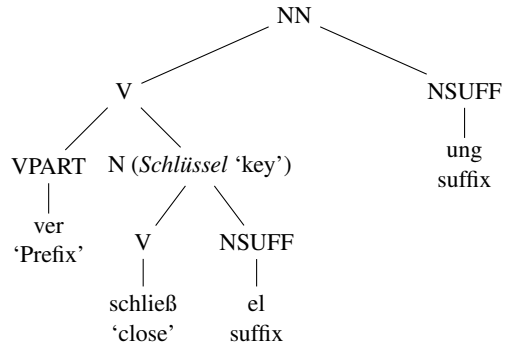

Figure 5: Morphological analysis of *Verschlüsselung*

Therefore, the frequencies should be augmented or replaced by ones from reliably lemmatized corpora. The derivational descriptions could be revised by an interested group of collaborators.

The refurbished CELEX database yields a sound basis for further developments in large-scale resources for morphology. We intend to use the lexicon and its frequencies to analyse morph boundaries within complex German words to analyse complex word structures, as done previously in Steiner and Ruppenhofer (2015). Other applications could be the development of a gold standard for morphological analyzers and parsers or as an training data for statistical approaches to hierarchical morphological parsing. So far, gold standards for German morphology as used by Cap (2014) and Steiner and Ruppenhofer (2015) provide only chains of morphs.

Last but not least, some parts of the script can be used for other purposes, especially for the transfer of existing German corpora from old to new spelling or the normalization of newer corpora with inconsistent spelling.

Acknowledgements

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

7. Bibliographical References

Bollmann, M., Krasselt, J., and Petran, F. (2012). Manual and semi-automatic normalization of historical spelling – case studies from early new high German. In *In Proceedings of KONVENS 2012 (LThist 2012 workshop*, pages 342–350.

Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., and Kokkinakis, D. (2009). Thinking Green:

Toward Swedish FrameNet++. In *FrameNet Masterclass and Workshop*.

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.

Dudenredaktion. (2013). *Duden - die deutsche Rechtschreibung*. Der Duden / in zwölf Bänden; das Standardwerk zur deutschen Sprache 1. Dudenverlag, Berlin, 26 edition.

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.

Gulikers, L., Rattink, G., and Piepenbrock, R. (1995). German Linguistic Guide. In Harald Baayen, et al., editors, *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, Philadelphia, PA.

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*.

Hanrieder, G. (1991). Robustes Wortparsing. Lexikonbasierte morphologische Analyse (komplexer) deutscher Wortformen. Master's thesis, Universität Trier.

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.

Sagot, B. (2014). DeLex, a freely-available, large-scale and linguistically grounded morphological lexicon for German. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

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)*.

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, University of Duisburg-Essen*, pages 49–57.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (2015). A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 674–680.

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.

8. 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.