Using Descriptive Generalisations
in the Acquisition of Lexical Data for Word Formation

Ulrich Heid, Bettina Säuberlich, Arne Fitschen

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Azenbergstraße 12, 70174 Stuttgart, Germany
{heid,tina,fitschen}@ims.uni-stuttgart.de

Abstract

This paper presents a method for acquiring data for a word formation analyser. There are several approaches to the analysis of complex words in German. As all of them have theoretical and/or practical drawbacks, we opt for a different approach: Instead of using linking elements, we make use of three different stem types, simplex, derivational, and compounding stems. Candidates for these can be generated automatically using knowledge about linguistic processes in German word formation. Based on the analysis of only a few phenomena we have gathered about 14.000 stems in a short time frame, all of them manually checked. As a result, certain wrong analyses can be avoided and ambiguities can be solved.

1. The Problems

In this paper, we describe an approach to the efficient semi-automatic acquisition of morphological data of high quality, which we use to enhance the precision and coverage of an analyser of German word formation.

This work is part of a research and development effort devoted to the creation of an analyser for complex words (rule system and lexicon). We intend to show how the choice of appropriate linguistic generalisations, beyond its usefulness for linguistic modelling, also supports an efficient acquisition strategy.

In the remainder of this section we first sketch the architecture of the analyser for complex words (section 1.1.) and compare it with other, similar tools (section 1.3.), before we describe the linguistic problem underlying the acquisition task. In section 2. we discuss the linguistic generalisations we use, and in section 3. we describe the acquisition scenario, the tools and procedures developed. Finally, section 4. is devoted to the results obtained so far, and in section 5. we discuss avenues for further development.

1.1. Context: DeKo, an Analyser for complex Words

Many word formation processes of German are productive. We understand productivity, here, in the sense of (Baayen 1992; Baayen 2001): using existing stems and existing word formation rules, speakers can build new complex words (often without being conscious at all of their novelty), and hearers can instantly analyse and understand them. Complex words created via productive word formation processes usually have a fully compositional (predictable) meaning, syntax, etc.

Examples include the derived adjective *fischhaft* ("fishlike") next to the more frequent items *katzenhaft, affenhaft* ("cat-like", "ape-like") etc.; or the compound *Kinderbecher* ("children's mug") next to *Kinderteller* ("children's plate"). Nominal compounds alone account for about two thirds of the words of a news corpus (cf. (Langer 1998)). Ten percent of the nouns in German newspaper text end in one of the suffixes *-ung, -heit, -keit, -schaft, -(i)tät, -ion*, which are part of productive derivation.

An analyser for complex words must be able to deal, at least morphologically, with such complex words created via productive processes. The DeKo[1] system has been designed for this purpose.

1.2. Resources for analysing complex Words

Within the task of constructing an analyser able to handle productively formed complex words, two subtasks concern lexicon building:

- A lexicon of simplex items must be available, supporting the analysis of the complex items by making base elements available;

- Since German nouns, verbs and (to a minor extent) adjectives vary in form when they occur as elements of complex words, ways must be found to capture this variation and to relate the variant forms to the corresponding lemmas.

Thus, we are confronted with a modeling problem and an acquisition problem at the same time.

Processes like Umlautung of the stem vowel (if it is |a|, |o|, |u|, or the diphtong |au|), the elision of schwa, the insertion of linking elements (e.g. *-s-, -er-, -en-,* etc.) or combinations of these make it a non-trivial task to divide complex words into their basic (simplex) components.

Compare the compounds *Haus·tür* ("front door") and *Häuser·meer* (lit.: "sea of houses"): In *Haus·tür*, the simplex stems *Haus* ("house") and *Tür* ("door") are concatenated without any linking element (we use · to mark the boundary), while in *Häuser·meer*, the linking element *-er-* is used, and furthermore the vowel |a| is changed to the corresponding umlaut. In the diminuitive form *Blüm·chen* ("little flower") of the noun *Blume* ("flower"), derivation involves the elision of the final schwa, and again Umlautung. These processes occur frequently in derivation and compounding (for details see section 4.).

---

[1] *DeKo* stands for "*De*rivations- und *Ko*mpositionsmorphologie des Deutschen". For a general overview of DeKo, see (Schmid et al. 2001). So far, the tool can not yet be tested online.

1.3. Approaches to the Analysis of complex Words

There are different ways to deal with German compounding, and some of them have been implemented.

(a) Complex words found in very large corpora may be listed, and the system only retrieves those items (and their analyses) which are contained in its lexicon ("lexicon approach"): this approach deliberately ignores productively constructed new items, providing, however, the possibility to produce highest quality output (manually corrected before admitted to the lexicon).

(b) Linking elements, elision and modification operations may be listed, and a rule system may be designed to apply them, possibly in combination. When a previously unseen complex word is encountered, the rules are applied until possible base forms are found ("rule approach").

(c) Wherever regular, knowledge about linking elements, as well as about elision and Umlautung may be noted together with paradigms of bases or with affixes ("subregularity approach"). For example, one may note that *-ung* takes *-s-* as a linking element when showing up as non-head of a compound, or that the affix *-lich* tends in most cases to enforce Umlautung (*blau/bläulich*, "blue/bluish").

Table 1 contains analyses of a few complex words, including productively constructed ones which are most likely not attested in corpora (they have English paraphrases in quotes). The analyses come from Gertwol[2], Word Manager[3], and DeKo. They allow us to draw conclusions as to which approach is followed by the systems.

The examples in (1a-d) in Table 1 test for compounds and their linking elements[4]. (1b) is not attested but possible, whereas (1d) is impossible, because *Hund* has never been found with *-es-* as a linking element (thus the form is not analysed by DeKo). If Lingsoft analyses (1d) as if it were correct, this may be evidence for the rule approach. The examples in (2), (3) and (4) concern derivation, with the (a)-type examples attested and the others being possible words. The unavailability of analyses for (2b), (4b, c), as well as for (1b, c) in WordManager seems to indicate quite clearly the lexicon approach[5].

If the lexicon approach is of limited use for the analysis of productively constructed words, the rule approach, on the other hand, is likely to accept too many, and possibly wrong items (see example 1d) or to lead to wrong analyses (e.g. *Bau·erhaltung*, "building maintenance" would receive a second, wrong, analysis as *Bauer·haltung* ("farmer's posture")). A merely rule-based approach would assign the compound *Vergnügungs·tempel* ("amusement hall") two analyses: *"Vergnügung* ("amusement") + *s* + *Tempel* (lit: temple, here: hall)", and *"Vergnügung + Stempel* (rubber stamp)", without any linking element.

The subregularity approach, however, is rather complex and unhandy for both modeling and acquisition, since it relies on subregularities for both stems (and their paradigms) and affixes, both of which have numerous exceptions. *-lich*, for example, is often a trigger for Umlautung (*blau/bläulich*, see above), but not always: *Frau/fraulich* ("woman/feminine, womanly") without Umlaut, but *Jungfrau/jungfräulich/*jungfraulich* ("virgin"/"virginal") obligatory with Umlaut.

Instead of the three approaches mentioned above, we have opted for another one, based on Fuhrhop (1998)'s notion of different types of stems.

2. Accounting for Stem Variation in German

The problems of lexical modelling of the formal variation in German nouns, verbs and adjectives, depending on their use as simplex items or in derivation or compounding seem to be best captured by the notion of stem types introduced by Fuhrhop (1998). Following her work, we do not assume the existence of (theoretically not well motivated) linking elements, but we assume that there are different types of stems: next to the simplex stem, there may be separate derivational and/or compounding stems.

2.1. Three Kinds of Stems

The variation between simplex, derivational and compounding stem is encountered with free morphemes as well as with bound morphemes (e.g. affixes, such as *-ung*, have a form *-ungs-* when they appear as a non-head element of a compound, see Table 5 below for details). Table 2 shows simplex, derivational and compounding stems, along with examples.

As Table 2 shows, some items (e.g. *Hund*), but not all have different forms for the three types of stems (e.g. *blau*). These stems must be listed in the lexicon of the analyser.

This approach does not provide an explanation for the fact that there may be several derivational or compounding stems for one lexeme: *Kind* has *Kinds-*, *Kinder-* and *Kindes-*, as in *Kinds·tod* ("cot death"), *Kinder·garten*, *Kindes·entführung* ("child kidnapping"). We classify all three stems as compounding stems (the derivational stem is *Kind-*, cf. *kind·lich*, *kind·isch*, *Kind·chen* ("childlike", "childish", "little child")). This procedure is sufficient for the purpose of analysis[6].

No.	Candidate word	EN paraphrase	Lingsoft	WordMngr	DeKo
1a	*Hundekuchen*	dog biscuit	+	+	+
b	*Hundetorte*	"dog cake"	+	−	+
c	*Hundenapf*	dog's bowl	+	−	+
d	*Hundesnapf	"dog's bowl"	+	−	−
2a	*Studentenschaft*	student body	+	+	+
b	*Maurerschaft*	"bricklayers' body"	−	−	+
3a	*katzenhaft*	cat-like	+	+	+
b	*fischhaft*	"fish-like"	−	+	+
4a	*lesbar*	readable	+	+	+
b	*schreibbar*	"writeable"	−	−	+
c	*schmeckbar*	"tasteable"	+	−	+

Table 1: Complex words and their analyses in Gertwol, WordManager and DeKo.

Word formation type	Category	free/ bound	Simplex stem	Comp./Der. stem	Examples	EN paraphrase
Derivation	noun	free	Hund	Hünd-	*hündisch*	servile
		bound	-tion	-tiön-	*Informatiönchen*	"little piece of information"
	verb	free	send(en)	send-	*Sendung*	broadcasting
	adj.	free	blau	bläu-	*bläulich*	bluish
Compounding	noun	free	Hund	Hunde-	*Hundekuchen*	dog biscuit
		bound	-tion	-tions	*Informationssperre*	news blackout
	verb	free	send(en)	sende-	*Sendezeit*	broadcasting time
	adj.	free	blau	blau	*blaugrau*	blue-grey

Table 2: Example data for simplex stem, derivational and compounding stem

2.2. Selectional Properties of Affixes

It is a well established fact that derivational affixes select bases by category: e.g. *-bar* ("-able") combines productively only with (transitive) verbs.

But there are other properties which play a role in the selection behaviour of affixes. One of them is the stem type: the affix *-lich*, for example, combines with the derivational stem, the same way as *-lein*, *-chen* and other affixes do. A few derivational affixes, however, select the compounding stem, the same one as used in compound formation. Examples are *-haft*, *-los* or the "semi-affixes" (*-frei*, *-freudig*, …): *Bauer/bauern·haft, bauern·los* ("farmer"/"farmer-like", "free of farmers"), with the same stem as *Bauern·haus* ("farmhouse"). For each derivational affix, the stem type which it selects can be identified (prominent examples are given in Table 3).

Some affixes in addition select only native (e.g. *-lich*) or only neoclassical stems (e.g. *-abel*), whereas others take bases of both types (e.g. *-isch*).

3. Acquisition of Stem Data

In the perspective of a stem-oriented approach to the analysis of complex words, which follows the descriptive work of Fuhrhop (1998), the acquisition problem is reduced to the task of finding, for each noun, verb and adjective lemma, next to the simplex stem also the derivational and compounding stem(s).

Selected category	Selected stem type	Affix expl	Examples (base/complex word)
noun	deriv.	$-chen_N$	*Bauer/Bäuerchen*
		$-lein_N$	*Katze/Kätzlein*
noun	comp.	$-schaft_N$	*Student/Studentenschaft*
		$-tum_N$	*Bauer/Bauerntum*
adjective	deriv.	$-lich_A$	*rot/rötlich*
		$-eln_V$	*schwach/schwächeln*
adjective	comp.	$-heit_N$	*schwach/Schwachheit*
verb	deriv.	$-lich_A$	*kaufen/käuflich*
	comp.	$-willig_A$	*kaufen/kaufwillig*

Table 3: Examples of affixes classified by their selection in terms of category and stem type of the base

3.1. Architecture

The principle of the acquisition architecture is quite simple. The procedures rely on the existing linguistic resources, i.e., a dictionary of about 20.000 simplex lemmas[7] and knowledge about morphophonological restrictions for affixes. The tool suite is schematized in Figure 1.

Candidates in our acquisition experiments come from a list of 2.14 million word form types extracted from corpora of news texts totalling about 200 million word forms.

---

[7]The lexicon contains in total about 45.000 lemmas of which a large portion are complex items which are lexicalised and therefore listed. For more information on the lexicon see (Lüdeling/Fitschen 2002).
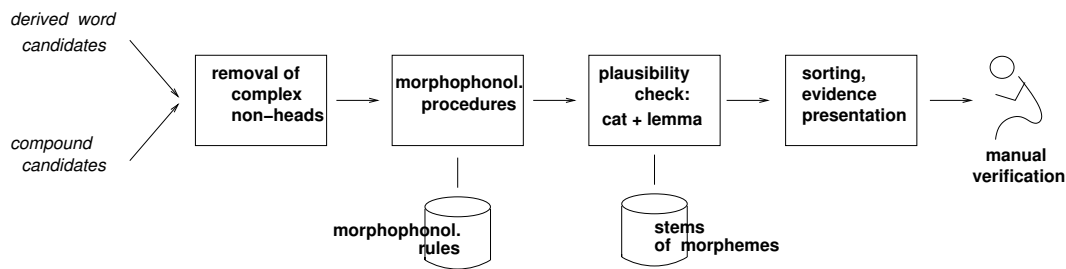
Figure 1: Tools used for the acquisition of compounding and derivation stems: use of the standard knowledge sources of the DeKo system: simplex stem lexicon and morphophonological rules

Derived words are grouped by suffix (e.g. all words ending in *-lein* together), compounds are processed by a rough decompounding heuristic, to remove compounds with complex non-heads from the candidate list[8]. Thereafter, the candidates are split into head and non-head, and non-heads are analysed morphophonologically (e.g. to "undo" Umlautung), in order to relate them to known simplex stems. This provides tuples of the following:

- corpus word analysed;

- form of the non-head found in the corpus word (i.e., derivation or compounding stem);

- candidate lemma of the non-head as derived by morphophonological rules and cross-checked for category and lemma status in the simplex dictionary.

3.2. Presentation for Manual Checking

The data produced by means of the tool suite shown above are rearranged for manual checking: the material is grouped by lemmas (cf. the two lines for *Auto* ("car"), from derived forms with *-chen* and *-lein*) and the person in charge of checking only needs to mark wrong stem hypotheses, as shown in Figure 2.

For manual checking of compounding stems, an Emacs mode has been designed. In the upper half of a split screen, the potential stems are displayed with the corpus frequency of their use as non-head in the 200 million word corpus (see Figure 3 for two sample lines with markings[9]). The corresponding examples are displayed underneath. For each stem, there are four actions: mark as "correct", mark as "wrong", or go to next or previous stem. Thus, a large candidate list can be processed in a fast, easy and accurate way.

30.214 stem candidates had to be manually checked. From these, 12.350 have been marked as valid non-heads, whereas 17.864 forms cannot function as compounding stems.

---

[8]The assumption is that compounds behave the same way as their bases with respect to compounding and derivation stems: *Buch* – *Büchlein* ("book", "little book") and *Sparbuch* – *Sparbüchlein* ("savings book" etc.) behave the same way, so *Sparbuch* can be removed (as redundant) from the candidate list for *buch/büch*; we are aware that there exist a few exceptions to this heuristic, as *Frau* – *fraulich* vs. *Jungfrau* – *jungfräulich*, *jungfraulich* indicates.

[9]The annotation "[p]" is provided by the tool to indicate paradigmic forms (plural, genitive, etc.).

4. Results

4.1. Overview

We have so far identified about 12.350 compounding stem forms for about 10.000 different noun lemmas, and about 1.800 derivation stem forms. The results were obtained with little effort. More corpora, including terminological ones, will be processed soon.

For example, for the noun *Gott*, ("god") we have three different compounding stems:

- *Gott-* (as in *Gott·vater* ("godfather")),

- *Gottes-* (as in *Gottes·furcht* ("godliness")),

- *Götter-* (as in *Götter·speise* ("ambrosia")).

In addition, the lexical entry for *Gott* includes two derivation stems:

- *Gott-* (as in *Gott·heit* ("deity")),

- *Gött-* (as in *gött·lich* ("divine")).

All stems proposed by the automatic acquisition system have been manually checked; derivational stems are hypothesized with very high accuracy, and the compounding stem candidates are of acceptable quality as well.

In many cases, we get multiple evidence for one stem hypothesis: for example, the derivation stem *schwäch-* for the adjective *schwach* ("weak") is present in $schwächeln_V$ ("be somewhat weak"), $schwächen_V$ ("weaken"), $Schwäche_N$ ("weakness"), $schwächlich_A$ ("somewhat weak"). When we have several word types which can be seen as converging evidence for one stem hypothesis, but no evidence of another hypothesis, as in the cases of *Vergnügungstempel* or *#Hundes-* discussed above, we have at least good reasons for excluding the non-attested forms.

4.2. Improving Disambiguation Power

On the basis of stem data (and knowledge about the morphological structure of the word in question), false ambiguities at the level of compound components can be solved easily; more examples of the kind of *Vergnügungstempel* are given in Table 4, below. Note that the heads under hypothesis 2 in the table, i.e. *Teller* ("plate"), *Teig* ("dough"), *Samt* ("velvet"), *Tat* ("act"), *Haltung* ("posture"), *Tand* ("trash") and *Aal* ("eel"), are all well-formed

```
Corpus word     Proposed analyses                                     Lemma hyp.  Deriv. Stem
------------------------------------------------------------------------------------------------
Aufständchen    Aufstand[nomen][masc][++]chen[suffix]                 Aufstand    Aufständ
Aufsätzchen     Aufsatz[nomen][masc][++]chen[suffix]                  Aufsatz     Aufsätz
Auftrittchen    Auftritt[nomen][masc][++]chen[suffix]                 Auftritt    Auftritt
Ausflügchen     Ausflug[nomen][masc][++]chen[suffix]                  Ausflug     Ausflüg
Auspüffchen     Auspuff[nomen][masc][++]chen[suffix]                  Auspuff     Auspüff
Autochen        Auto[nomen][neut][++]chen[suffix]                     Auto        Auto
Autolein        Auto[nomen][neut][++]lein[suffix]                     Auto        Auto
Baggerlein      Bagger[nomen][masc][++]lein[suffix]                   Bagger      Bagger
```

Figure 2: Examples of the input to manual correction (left) and the (processed) result: derivation stems of nouns
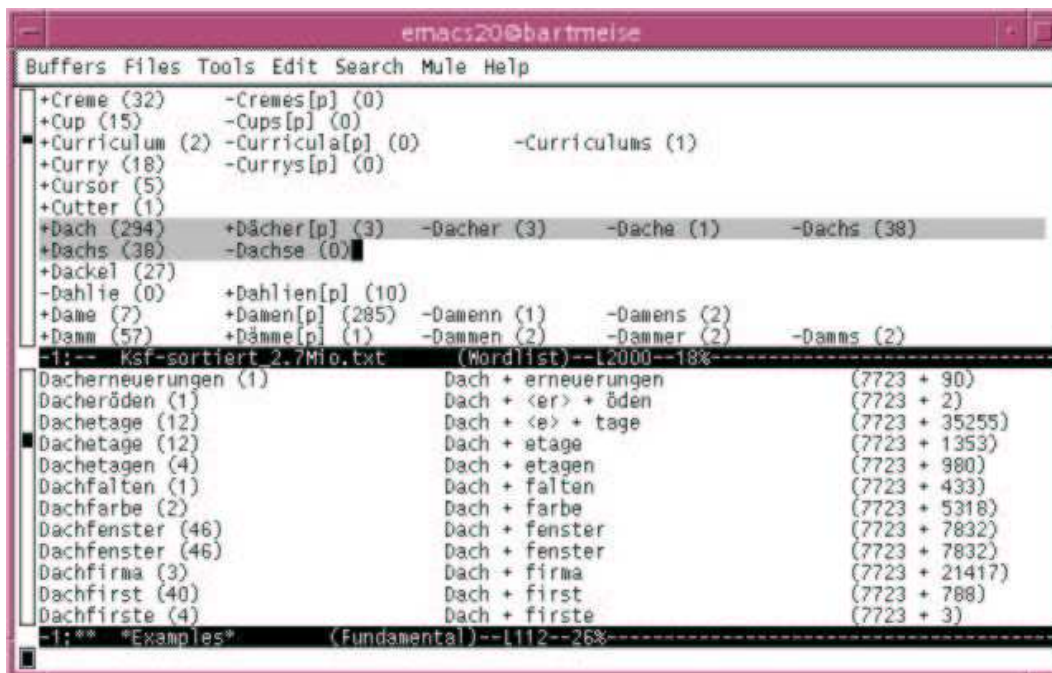

Figure 3: Screen shot of the Emacs mode for manual correction of compounding stems, with markings

nouns; however, the compound stem hypotheses for examples (1) through (5) under hypothesis 2 are wrong.

Table 4 also contains examples of cases where the ambiguity cannot be solved with our approach (examples (6) to (8), marked with "(+)" on the wrong but possible analysis). In #*Antrags·teller*, the compound stem hypothesis is not wrong, as *Antrags·formular ("application form"), Antrags·frist* ("deadline for application") show. Similarly, *Generals-* is found in *Generals·uniform ("general's uniform"), Generals·besprechung* ("general's conference"). In the case of *Kursaal, Kurs-* can indeed not be a compound stem of *Kur* ("cure"), but it is a compound stem of *Kurs* ("rate"). To decide that the second hypothesis for the analysis of the examples (6) to (8) in Table 4 is unacceptable, we need world knowledge rather than linguistic knowledge.

We have extracted noun+noun compounds from a 200 million word corpus with ambiguous analyses in the above-mentioned sense. Among the top 150 ambiguous items, sorted according to their frequency of occurrence in the corpus (these items have a total frequency of 72.795 in 200 million), only 17 ambiguities could not be resolved with the methods described here; thus 88,67 % of the types from

this small sample are analysed correctly.

4.3. Improving Generalisation Power

Moreover, we can generalise over large portions of the complex vocabulary, by formulating entries for the derivational and compounding stems of affixes: the stems are inherited from the affixes (i.e. the derivational heads) to the derived words. Examples of a few frequent affixes are given in Table 5, below, along with the type frequency of the compounds where they occur as non-heads, as well as with the token frequency of the respective compounds, each within our 200 million words news corpus. In addition, we give one example of each type.

These types alone make up for 14 % of all noun+noun compounds in our corpus.

5. Conclusion

We discussed a tool suite for the semi-automatic acquisition of morphological data for an analyser of complex words. The approach taken is based on Fuhrhop (1998)'s distinction of simplex stems, derivational stems and compounding stems. The tools provide hypotheses of high qual-

#	Lemma (paraphrase)	Hypothesis 1	Status	Hypothesis 2	Status
1	Schriftsteller (writer)	Schrift·steller	+	Schrift·s·teller	−
2	Bahnsteig (platform)	Bahn·steig	+	Bahn·s·teig	−
3	Wohnungsamt ("housing office")	Wohnung·s·amt	+	Wohnung·samt	−
4	Kulturetat (culture budget)	Kultur·etat	+	Kultur·e·tat	−
5	Machterhaltung ("preservation of power")	Macht·erhaltung	+	Macht·er·haltung	−
6	Antragsteller (proposer)	Antrag·steller	+	Antrag·s·teller	(+)
7	Generalstand ("general's position")	General·stand	+	General·s·tand	(+)
8	Kursaal ("hall in a spa")	Kur·saal	+	Kurs·aal	(+)

Table 4: Examples of ambiguities at the level of compound component identification

Simplex stem	Type f	Token f	Derivation stem	Examples	Compounding stem	Examples
-ung	36.876	412.993	—	—	-ungs-	Bildungslücke
-heit	2.888	43.556	-heit-	einheitlich	-heits-	Einheitsessen
-(t)ion	12.586	108.691	-(t)iön-	Funktiönchen	-(t)ions-	Funktionsumfang
-ist	1.651	8.452	-ist-	linguistisch	-isten-	Linguistenarbeit

Table 5: Simplex, derivation and compounding stems of frequent derivational affixes

ity and make the manual checking relatively easy and efficient.

The approach is linguistically sound, in that it does not require us to assume the existence of linking elements the linguistic status of which is not very clear (are they morphemes, do they have a category, do they have selectional properties, etc.). In comparison with the rule approach (cf. section 1.3.), we expect that our approach will considerably reduce the number of ambiguities in the analysis of complex items (cf. Table 4).

It should be noted, however, that the rule approach and the subregularity approach produce at least reasonable results on a large portion of German compounds. From 21.424 compounds analysed with our tools, 43.22 % have a compounding stem which is identical to the simplex stem. Another 54.89 % have one of the nine linking elements -*s*-, -*n*-, -*en*-, -*e*-, -*er*-, -*es*-, -*o*-, -*nen*-, -*ns*-, -*ien*-, -*ens*-, and only less than 2% have elision and/or umlaut, possibly with linking elements (there is considerably more Umlautung in derivation). Thus, for a rough first analysis, a less sophisticated approach may be sufficient.

Our approach has, however, additional potential with respect to the analysis of neoclassical word formation. In line with the stem approach and with Lüdeling et al. (2002), we also assume the existence of two (or more) stems (called 'short' and 'long', for the moment) in neoclassical word formation. Thus, we note the stems *demonstr*- and *demonstrat*-, to cover *demonstr·ieren, Demonstr·ant, demonstr·abel* vs. *Demonstrat·ion, demonstrat·iv*. Such stems can again be acquired with the same principles and tools as shown, because of the selection preferences of neoclassical affixes. Acquiring a sizeable lexicon of stems for neoclassical word formation will be the next major application of the tool suite described here.

6. Acknowledgements

7. References

Baayen, Harald (1991). "Quantitative Aspects of Morphological Productivity." In G. Booij and J. v. Marle, eds., *Yearbook of Morphology 2*, 109 – 150. Foris, Dordrecht.

Baayen, Harald (2001). *Word Frequency Distributions.* Kluwer Academic Publishers, Dordrecht, 2001.

Domenig, Marc and ten Hacken, Pius (1992). *Word Manager: A system for Morphological Dictionaries.* Olms, Hildesheim.

Fuhrhop, Nanna (1998). *Grenzfälle morphologischer Einheiten.* Stauffenburg-Verlag, Tübingen.

Hausser, R. (1996). *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994.* Max Niemeyer Verlag, Tübingen, 1996

Koskenniemi, K. and Haapalainen, M.: "GERTWOL – Lingsoft Oy", in (Hausser 1996): 121-140.

Langer, Stefan: "Zur Morphologie und Semantik von Nominalkomposita", in: (Schröder et al. (Ed.) 1998), 83-96

Lüdeling, Anke and Schmid, Tanja and Kiokpasoglou, Sawwas: "Neoclassical word formation in German", in: *Yearbook of Morphology 2001*, to appear 2002

Lüdeling, Anke and Fitschen, Arne: "An integrated lexicon

for the automatic analysis of complex words", to appear in: *Proceedings of Euralex 2002*, Copenhagen, 2002

Schmid, Tanja and Lüdeling, Anke and Säuberlich, Bettina and Heid, Ulrich and Möbius, Bernd: "DeKo: Ein System zur Analyse komplexer Wörter", in: Henning Lobin (Ed.): *Sprach- und Texttechnologie in digitalen Medien* - Proceedings der GLDV-Frühjahrstagung 2001 (Gießen), 49-57; see also:
`http://www.uni-giessen.de/fb09/ascl/gldv2001`

Schröder, Bernhard and Lenders, Winfried and Hess, Wolfgang and Portele, Thomas (Ed.): *Computer, Linguistik und Phonetik zwischen Sprache und Sprechen – Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache – KONVENS-98*, (Frankfurt: Lang) 1998