An Integrated Lexicon for the Automatic Analysis of Complex Words

Anke Lüdeling
Institut für Kognitionswissenschaft
Universität Osnabrück
Katharinenstr. 24, 49069 Osnabrück, Germany
aluedeli@uos.de

Arne Fitschen
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Azenbergstraße 12, 70174 Stuttgart, Germany
fitschen@ims.uni-stuttgart.de

Abstract

In this paper, we motivate the need for a large-scale computational lexicon that is used as a resource for a morphological word analysis tool in German. Complex words account for more than two thirds of the vocabulary of German, which makes their unambiguous analysis crucial for all kinds of linguistic processing. The important linguistic features that are needed are introduced, and an implementation of the lexical resource is presented. First linguistic features that are needed for word analysis are introduced, then an implementation of the lexicon is presented, and finally we sketch the interaction between the lexicon and the word analysis tool.

Introduction

Why do we need a large-scale computational lexicon for the automatic analysis of word formation? As in many languages, word formation in German is highly productive. According to Ortner & Müller-Bollhagen [1991] (p. 3), about two thirds of the vocabulary are nominal compounds alone.[1] The median of lemma frequencies in any given text is 1. Thus, more than 50% of lemma types occur only once in the text, regardless of text length. This shows that it is not possible to work with a finite lexicon for any purpose that needs to analyze unrestricted text [Baayen 1992, 2001; Evert & Lüdeling 2001]. Therefore, for many computational linguistic applications an analysis component is necessary. This is possible because productively formed words tend to be morphosyntactically and semantically regular.[2] Word formation is, on the other hand, restricted on many linguistic levels. In order to express these restrictions one needs to have a lexicon that lists the word formation elements (morphemes, if you wish) together with all the relevant information. Although this is not controversial in theoretical and descriptive linguistics, we are not aware of any large-scale computational word-formation component that refers to such a lexicon. Some of the existing large computational lexicons focus on semantic and conceptual information (WordNet[3]) or on phonological information [Portele et al. 1995] because they are built for different purposes. Even lexicons that are built as a resource for word formation components contain only some of the relevant types of information ([CISLEX 1995; CELEX 2001; Quasthoff 1995]). It is also not possible to simply gather the information from theoretical or

empirical works on word formation since they typically provide only a few examples for each phenomenon.
In this paper we (1) discuss how a large-scale computational lexicon that supports the automatic analysis of complex words should look like and (2) introduce one implementation of such a lexicon for German.
The paper is organized as follows: First we motivate the need for a complex lexicon linguistically by giving some examples for restrictions on word formation in German. Then we describe the lexicon that we have constructed in order to cope with these requirements before giving some implementation details. Finally, we show how the lexicon and a word analysis component interact.

Restrictions on Word Formation in German

Word formation is restricted on all linguistic levels. In this section we first give a few examples for the kinds of linguistic categories that restrict word formation (see, for example, [Fleischer & Barz 1992; Ortner & Müller-Bollhagen 1991; Kühnhold et al. 1978]) in order to motivate the need for a lexicon that incorporates information on all these categories. Then we speak about stem changes like umlauts, elision and linking elements which make German word formation especially difficult.
Restrictions on word formation include the following:

part of speech category the adjective forming suffix *-bar*[4] attaches productively to verb stems: *annehmbar* ``acceptable" from *annehmen* "to accept".

argument structure *-bar* productively only attaches to transitive verbs: **schlafbar* ``sleepable" from intransitive *schlafen* "to sleep".

origin In many Germanic languages we find morphological elements of Latin/Greek/Romance origin alongside the native elements. These so-called neoclassical elements sometimes behave differently from the native elements in word formation: the adjective forming suffix *-abel*, for example, combines only with neoclassical elements: *akzeptabel* "acceptable" from *akzeptieren* "to accept", but not **annehmabel* from *annehmen* "to accept". *-bar*, on the other hand, combines with neoclassical and native elements: *akzeptierbar*, *annehmbar*.

phonology The noun forming suffix *-ei* attaches to trochaic words ending in a schwa-syllable: *Bäckerei* "bakery" from *Bäcker*. If *-ei* wants to attach to a noun that does not end in a schwa-syllable, the allomorph *-erei* is used: *Spielerei* "playing around" from *Spiel* "game". The stress structure of nouns for example provides cues to the origin (and thus to the morphological combinability). Two-syllable native words tend to be trochaic (*Hase* "rabbit", *Blume* "flower"), while two-syllable foreign words tend to be iambic (*Student* "student", *Konsens* "consensus").

semantics Different kinds of semantic information such as the mass/count distinction or conceptual classes restrict word formation. Nouns denoting time spans, for example, combine with the adjective forming suffix *-lich*, as in *Stunde* "hour" – *stündlich* "hourly", *Tag* "day" – *täglich* "daily", etc.

In order to have a maximally constrained word-formation component we thus need to formulate rules that can refer to all of the relevant categories. In addition to such rules we have to deal with stem changes in German.

Morphological elements sometimes change their form in word-formation in German – German uses linking elements, elision and umlauts. These changes are not always regular: *Bund* "union" and *Grund* "reason, basis", for example, which have the same inflectional class, behave differently in compounding:[5] *Bundes·gesetz* "federal law" vs. *Grund·gesetz* "constitutional law". *Bund* always becomes *bundes* when it is the non-head in a compound while *Grund* always is *grund*. Some nouns have more than one form in compounding.

We find similar changes in derivation. *Frau* "woman", for example, which is often *frauen* in compounding (as in *Frauen·fußball* "women's soccer") has two forms in derivation, *frau* in *frau·lich* "womanly" and *fräu* in *Fräu·lein* "Miss". Finally, the final *e* in *Sprache* "language" is elided in derivation: *sprach·lich* "linguistic".

From these examples it becomes evident that we cannot use a rule-based mechanism for dealing with linking elements, umlauts and elision (see also [Fuhrhop 1998; Krott 2001]). Rather, we have to list the possible forms in the lexicon. Following Fuhrhop [1998] and Eisenberg [1998] we call these forms compounding stem forms and derivation stem forms. Later in this paper, we will see examples where the use of such stem forms drastically reduces the number of ambiguities in automatic analysis.

The lexicon

A word formation component that automatically analyzes complex words should be able to refer to all the relevant information discussed above in order to maximally reduce ambiguities. The prerequisite is a very detailed lexicon where all the information is given for each morphological entity. We have developed the lexicon described in this paper (henceforth: the Integrated Lexicon) to be used with a specific word formation component[6] but it can, in principle, be used with any word-formation component.

Content

The Integrated Lexicon is designed to contain information from all linguistic levels. It builds on a lexicon that was designed as an inflectional lexicon, see [Schiller 1994; Lezius et al. 2000]. Elements are categorized by part-of-speech. For each part-of-speech category there are relevant features, like *gender* for nouns, *argument structure* for verbs (semi-automatically acquired by Eckle-Kohler [1999]), or *gradation forms* for adjectives. In addition to these features, there are common features for all lexical entries.

The information on the values for these features has to be acquired semi-automatically, where some kinds of information are more difficult to acquire than others.

Some of the features are

citation form and stem The citation form is not necessarily the stem that is used in word formation. For verbs, German uses the infinitive as a citation form (*schlafen* "to sleep") while most word formation processes use the stem (*Schlafzimmer* "bedroom").

origin Here we state whether an element is native, neoclassical, English, etc. This is sometimes difficult to decide [Lüdeling et al. 2002; Lüdeling & Schmid 2002].

morphological status Here we state whether an element is free or bound, whether it is simplex or complex, whether it is a 'regular' word or an abbreviation etc.

selection Here we state whether an element selects other elements in word formation.[7] There must be at least one rule for each selecting element in the word formation component.

compounding stem form and derivation stem form Here we note all the different word formation stem forms of an element. The stem forms are acquired by using a simple analysis tool including a primitive mechanism for elision and linking that suggests a number of possibilities for each noun; the correct possibility has to be marked manually (see [Heid et al. 2001] for details).

phonological information The phonology of each element is given in SAMPA code, there is also information about the syllable structure, stress and final devoicing. The SAMPA encoding stems from the German text-to-speech system [Möbius 1999].

Above we stated that semantic information is also important to restrict word formation. We haven't yet included any kind of semantic information in the lexicon besides the annotation of different kinds of names.
Table 1 gives an overview over a few example entries, the free native nouns *Bund* "union", *Funken* "spark" and *Funk* "broadcast", the bound neoclassical noun *anthrop*, the free native verb *kosten* "to cost" and the bound nominal element *-ung*.

citation form	stem	morph. status	Selection	origin	derivation stem form	compounding stem form
Bund "union"	bund	free simplex	-	native	bünd	bundes
Funken "spark"	funke	free simplex	-	native	fünk	funken
Funk "radio"	funk	free simplex	-	native	funk	funk
Anthrop- "human"	anthrop	bound simplex	-	neoclassical	-	anthropo
kosten "to cost"	kost	free simplex	-	native	kost	kost
-ung	ung	bound simplex	+	native	-	ungs

Table 1: A few example entries (we left out part-of-speech information, phonological information and inflectional class as well as part-of-speech specific information). *-ung* is a nominalisation suffix which produces event nouns and result nouns from verbs.

Implementation

The Integrated Lexicon is represented as a set of XML files, roughly one per (major) part of speech. The XML formalism[8] allows for the well-defined structuring of all different kinds of information as well as for a flexible handling of changes to the structure. There are tools for

the maintenance of the resource as well as for the export of (parts of) the information contained.
The XML files were generated with Perl scripts from a previously developed lexical resource, a relational database that stored morphological and syntactic information (for details, cf. [Lezius et al. 2000]). The new resource is again stored in a relational database, but here, the data model was derived automatically from the definition of the lexicon's structure (the 'document type definition', DTD). In case of changes to the structure the database can be automatically rebuilt from the XML files.
A tool for adding new entries to the lexicon uses the DTD in a similar fashion: From the types of information defined in the DTD, textual information, or a range of values for an attribute, a graphical user interface (GUI) is derived which allows for editing the textual information or for selecting values. Thus, the user interface does not have to be adepted if there are changes in the resource's structure. There is, however, a trade-off between GUI design and flexibility in data handling: To ensure maximum flexibility, it is not possible e.g. to emphasise by formatting the representation of important elements, since the information can only be represented as defined in the DTD.
For the extraction of information from the lexicon, so called 'stylesheets' can be used, a standardized means of transforming XML documents into different formats. Parts of the lexicon can be selected, re-grouped, presented according to formatting parameters, and transformed easily.
The biggest advantage of using XML for the representation of the files is the fact that the syntax of the lexicon entries can be automatically checked for validity, that is its conformity to the structure defined in the DTD. Hence, it is impossible to insert wrong entries or unknown features in the lexicon. Furthermore, for the definition of feature values an enumeration type can be used: Again, typing errors will be detected by the XML parser. Of course this only holds for the features whose values can be enumerated.[9] Besides this, the ability to link information within one or more XML documents, and again to automatically check for inconsistencies, is very useful for a complex lexicon.
There are drawbacks to using XML, however. Processing large XML documents takes quite a while with current tools, and the memory requirements are high. This is one of the reasons (besides security reasons and multi-user accessibility) for continuing to use a database for storing the lexicon entries. Furthermore, unlike syntactiv validity, there's no means of checking semantic validity. So, if the DTD requests a citation form, an entry like `<citation_form id=""word1">Nosnense</citation_form>` is syntactically well-formed and will be parsed, but there is no tool to decide whether the element's content is correct.

Interaction between the Integrated Lexicon and a Word-Analysis Component

The Integrated Lexicon is used to restrict the word analysis rules in two ways: unspecific linking and elision rules that lead to incorrect answers can be avoided and rules that involve selecting elements can be written.
Consider a word formation component that only has stems and affixes and rules for linking elements, elision and umlauts at its disposal. Such rules would, for example, analyse a compound by starting at the end of a word and then processing letter by letter until a known word is found. Then the mechanism would try to find out whether the remaining part is also

a known word. If not, it would elide all possible linking elements, change umlauts etc. until the remaining part matches a known word. This is more or less how most analysis tools work [Lezius et al. 1998; Langer 1998]. This mechanism leads to unwanted ambiguities, however. Consider the compound *Vergnügungstempel* "amusement hall" which would have (at least) two possible analyses: *Vergnügung* "amusement" + *Tempel* "lit: temple, here: hall", where the linking element *s* is used, or *Vergnügung* + *Stempel* "rubber stamp" without any linking element. Morphologically the second possibility is incorrect (in addition to its semantic oddness), since words ending in the nominal suffix *-ung* always have the linker *s* in compounds. If the analysis component has access to a lexicon that includes compounding stem forms this mistake is avoided. In addition, the stem forms are often quite different from the corresponding word, so that an automatic analysis would be difficult or impossible: compare, for example, *Tafel* "table, board" with the diminuitive form, *Täf·lein* "tiny table", or the neoclassical noun *Insel* "island" which has a derivation stem form *insul*, as in *insular* "insular" or *Insulaner* "islander".

Rules involving selecting elements can only be written if the relevant information is available. Again, many ambiguities are avoided if such rules can be formulated. Consider the adjective forming suffix *-bar*, for example, which is homograph to the free noun *Bar* "bar, pub" and the free adjective *bar* "in cash". The suffix *-bar* attaches productively only to transitive verbs and the free adjective *bar* does not appear in complex words at all. If we have to analyze the word *Pianobar* "piano bar", we can therefore safely assume that this is a noun+noun compound.

Conclusion

We have motivated the need for a lexicon containing more kinds of information than those currently available in order to provide for a less ambiguous automatic analysis of complex German words. We have shown that the concept of compounding stem forms and derivation stem forms helps avoiding ambiguities that arise if we only have a stem lexicon and unspecific linking rules. In addition, information on all linguistic levels is crucial for maximally restricting word analysis rules. We have implemented a lexicon containing the information relevant for word formation and have shown how this lexicon interacts with a word analysis component.

At the moment we have 12000 compound stem forms for about 10000 different nouns (out of the 25000 nouns in the lexicon), and we got about 1000 derivation stem forms.

References

[Baayen 1992] Baayen, R. Harald, 1992. Quantitative aspects of morphological productivity, in *Yearbook of Morphology 1991*, Foris, Dordrecht, 109-150.

[Baayen 2001] Baayen, R. Harald, 2001. *Word Frequency Distributions*. Kluwer, Dordrecht

[CELEX 2001] Piepenbrock, Richard (CELEX), 2001. CELEX, the Dutch Centre for Lexical Information. [http://www.kun.nl/celex/].

[CISLEX 1996] CISLEX, 1996. CISLEX - das Wörterbuch am CIS. [http://www.cis.unimuenchen.de/projects/CISLEX.html].

[Quasthoff 1995] Quasthoff, Uwe, 1995. Wortschatz Deutsch. [http://wortschatz.uni-leipzig.de/].

[Eckle 1999] Eckle-Kohler, Judith, 1999. *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora.* Logos Verlag, Berlin.

[Eisenberg 1998] Eisenberg, Peter, 1998. *Grundriss der deutschen Grammatik. Band 1: Das Wort.* J.B. Metzler, Stuttgart.

[Evert & Lüdeling 2001] Evert, Stefan and Lüdeling, Anke, 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? in: *Proceedings of Corpus Linguistics 2001,* Lancaster.

[Fleischer & Barz 1992] Fleischer, Wolfgang and Barz, Irmhild, 1992. *Wortbildung der deutschen Gegenwartssprache.* Max Niemeyer Verlag, Tübingen.

[Fuhrhop 1998] Fuhrhop, Nanna, 1998. *Grenzfälle morphologischer Einheiten.* Stauffenburg-Verlag, Tübingen.

[Heid et al. 2001] Heid, Uli, Säuberlich, Bettina, and Fitschen, Arne, 2001. *Using descriptive generalizations in the acquisition of lexical data for a word formation analyzer.* Submitted for publication.

[Krott 2001] Krott, Andrea, 2001. *Analogy in Morphology. The Selection of Linking Elements in Dutch Compounds.* MPI Series in Psycholinguistics, Nijmegen.

[Kühnhold et al. 1978] Kühnhold, Ingeburg, Putzer, Oskar, and Wellmann, Hans, 1978. *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache 3: Das Adjektiv.* Pädagogischer Verlag Schwann, Düsseldorf.

[Langer 1998] Langer, Stefan, 1998. Zur Morphologie und Semantik von Nominalkomposita, in *Proceedings of the KONVENS 1998. Computers, linguistics and phonetics between language and speech.* Bonn.

[Lezius et al. 1998] Lezius, Wolfgang, Rapp, Reinhard, and Wettler, Manfred, 1998. A freely available morphological analyzer, disambiguator, and context sensitive lemmatizer for German, in *Proceedings of the COLING-ACL 1998,* 743-747.

[Lezius et al. 2000] Lezius, Wolfgang, Dipper, Stefanie, and Fitschen, Arne, 2000. IMSLex - representing morphological and syntactical information in a relational database, in U. Heid, S. Evert, E. Lehmann, and C. Rohrer, editors, *Proceedings of the 9th EURALEX International Congress*, Stuttgart, 133-139.

[Lüdeling & Schmid 2002] Lüdeling, Anke and Schmid, Tanja, 2001. Does origin determine the combinatory properties of morphological elements in German?, to appear in J. DeCesaris, editor, *Proceedings of the Third Mediterranean Meeting on Morphology*, Barcelona, 2001.

[Lüdeling et al. 2002] Lüdeling, Anke, Schmid, Tanja, and Kiokpasoglou, Sawwas, 2002. Neoclassical word formation in German, to appear in: *Yearbook of Morphology 2001.*

[Möbius 1999] Möbius, Bernd, 1999. The Bell Labs German text-to-speech system, in: *Computer Speech and Language*, 13, 319 - 357.

[Ortner & Müller-Bollhagen 1991] Ortner, Lorelies and Müller-Bollhagen, Elgin, 1991. *Deutsche Wortbildung. Substantivkomposita (Komposita und kompositionsähnliche Strukturen 1).* Walter de Gruyter, Berlin, New York.

[Portele et al. 1995] Portele, T., Krämer, J., and Stock, D., 1995. Symbolverarbeitung im Sprachsynthesesystem Hadifix, in: *Proceedings der 6. Konferenz Elektronische Sprachsignal-verarbeitung 1995*, Wolfenbüttel, 97-104

[Schiller 1994] Schiller, Anne, 1994. *Deutsche Flexions- und Kompositionsmorphologie auf 2-Ebenen-Basis.* Technical report, Institut für maschinelle Sprachverarbeitung.

[Schmid et al. 2001] Schmid, Tanja, Lüdeling, Anke, Säuberlich, Bettina, Heid, Ulrich, and Möbius, Bernd, 2001. DeKo: Ein System zur Analyse komplexer Wörter, in: H. Lobin, editor, *Proceedings der GLDVFrühjahrstagung 2001*, 49 - 57.

Endnotes

1 In this paper we are not dealing with inflection. Whenever we use the term 'complex word' we refer to word formation, in particular compounding and derivation. We discuss examples and an implementation for German, but similar lexicons are required for word formation components in other languages.

2 Regularity is, in fact, one of the defining properties of morphological productivity. Complex words

that are in any way irregular, have to be listed in a lexicon with internal structure, if desired. In the following we will focus on regularly formed words.

3 http://www.cogsci.princeton.edu/wn/

4 We mark bound morphemes with a'-'.

5 We mark morpheme boundaries with a where instructive.

6 The word formation component DeKo (for Derivation "derivation" and Komposition "compounding", financed by the State of Baden Württemberg) ran from Jan 2000 to June 2001. See also http://www.ims.uni-stuttgart.de/projekte/DeKo/ and [Schmid et al. 2001].

7 This does not necessarily coincide with boundness (there are free elements which select and bound elements which do not select, see [Lüdeling et al 2002] - therefore we speak about morphological status and selection rather than of affixes and stems.

8 See http://www.w3.org/xml for more information on XML .

9 At the moment it seems that 'Schemata' will supersede DTDs in representing the document's structure. Here, one can restrict data types, value ranges etc. (see http://www.w3.org/xml/Schema).