Word Formation in Computational Linguistics

Pius ten Hacken
Universität Basel & Anke Lüdeling Universität Osnabrück

Motivation: why do we need a word formation component for computational linguistic applications?

Where does word formation information play a role in computational linguistics? Think of information retrieval where the parts of a word might contain the important information. If we want to find information in German on relaxation techniques, we need to look for the verb *entspannen* 'to relax' as well as for the derived noun *Entspannung* 'relaxation' and compounds containing *Entspannung*, such as *Entspannungsantwort* 'relaxation response', *Tiefenentspannung* 'deep relaxation' or *Entspannungsübung* 'relaxation exercise' etc. Another example are Text-To-Speech systems where the structure of a word can tell us where the stress goes. Consider English words containing so-called neoclassical affixes (affixes of Latin or Greek etymology). Some of these affixes, such as *–ation* influence the stress of a word: *re'lax* vs. *relax'ation*. Even such seemingly 'basic' components as part-of-speech taggers can (and often do) use word formation information, if only as heuristics. If a tagger encounters an unknown English word ending in the letters <ous>, for example, it can guess that this must be an adjective.

Word formation information is thus important for many computational linguistics applications. But why do we need a word formation component? There are large machine-readable lexicons available today. Why can't we just use these? First, we have to distinguish between computational linguistic applications (by application we mean 'higher' systems like machine translation systems, text understanding systems etc. that include a number of components) that deal with a fixed text/set of texts (which can, of course, work with a finite lexicon) and applications that deal with unseen text.[1] These applications make use of basic components such as taggers, lemmatizers / stemmers, parsers etc. Often these components (and accordingly the application) – no matter

[1] This distinction does not primarily depend on the application type (e.g. machine translation, text summarization), but rather on such factors as the context of use and the strategy chosen to solve the problem at hand. Thus a machine translation system for weather forecasts such as Météo (with its restrictions of use as described by Chandioux (1989)) has a closed set vocabulary. If the domain of a machine translation system is less restricted it is bound to be confronted with new, unseen words. In the case of text summarization, a different distinction can be made (cf. Endres-Niggemeyer (1998)). Here there is one strategy which basically consists of determining a set of key words and ranking sentences of the text to be summarized in terms of their use of key words and their position in the text. A summary is then compiled by putting together the highest-ranked sentences. In this strategy, a fairly small vocabulary is sufficient, because no attempt to arrive at a full-scale analysis of the text is undertaken. Unseen words occur frequently, but are ignored. Of course a more sophisticated strategy which assumes a stage at which the structure and meaning of the text is represented as a basis for the summary, cannot ignore unseen words.

whether they deal with limited or unlimited text – use some kind of dictionary or lexical database where the necessary information about the words that are used – depending on the application this could be part-of-speech category, semantic information, the SAMPA code etc. – is stored.[2] Sometimes that information includes word formation information. However, if one deals with unseen text, there is a large chance of encountering words that are not listed in that dictionary, no matter how large the dictionary is – therefore one needs a strategy to deal with unseen words. In this tutorial we want to acquaint you with some of the problems of the automatic treatment of word formation.

The tutorial is organized as follows: first we give a little lexical statistics that shows how often new words occur in a text. Then we give some reasons for this – at the same time we also distinguish between different kinds of 'unseen' or 'new' words and explain which ones can be dealt with in a morphological component. Then we review a few word formation concepts and problems before spending some time to introduce two existing word formation systems, DeKo and Word Manager. We include an appendix with pointers to word formation systems in a number of languages with urls and/or references.

A little lexical statistics about unseen words in texts

In this section we want to illustrate that if you deal with unseen natural text, the probability of encountering unseen words is high. That means that a finite lexicon – no matter how large it is – can **never** be sufficient in dealing with unseen text. Here we concentrate on (very basic) statistics; the mechanisms that are used to produce new words are introduced in the next section.

The question we need to ask here is: how likely is it that we encounter a new word after we have sampled a certain amount of text? Or, phrased differently, how large would a corpus have to be so that it contains all the words of a language? We can count the types, that is, the different words[3], in a given text and calculate a so-called type-token ratio which tells us how often each type occurs (in other words: how many tokens of each type can be found). We can then order these types according to token frequency. Table 1 shows some lemma types and their frequencies from the *Stuttgarter-Zeitung* Corpus (a German newspaper corpus, about 36 million tokens). The most frequent type is the definite article which occurs roughly 3,5 million times. The second most frequent type is the comma which occurs about 1,8 million times etc. At the end of the table we see many, many types that occur only once – here we see the last nine types in an alphabetic order.

We can now count the number of types that occur once, twice, three times etc. (we then have the frequency of frequencies) – this is called a frequency spectrum, see Table 2. Here we see that roughly 404,000 types occur once, roughly 97,000 types occur twice etc. At the bottom left corner we see that there is one type that occurs 3,5 million times

[2] There are applications and components that deal with any kind of text purely statistically without ever referring to a lexicon. We will not talk about those here.

[3] For our purposes here it does not matter whether we count word-form types or lemma types; the results are qualitatively the same.

– that is, of course, the definite article. Table 2 shows a very typical picture: in any given text we find a few very frequent types and many rare types.[4]

type	frequency	type	frequency
d (definite article)	3,571,573	...	...
,	1,848,517	Zytomegalievirus 'zytomegalievirus'	1
.	1,605,763	Zytomir (geographic name)	1
ein (indefinite article)	710,719	Zytos 'cell'	1
und 'and'	708,531	zytotoxische 'toxic for cells'	1
in 'in'	613,876	Zywietz (last name)	1
PPER (personal pronoun)	536,174	Zyzik (last name)	1
sein 'to be'	534,056	ZZ-Top-Hit 'ZZ-Top hit'	1
"	408,708	ZZ-Top-Käfer-Nachbau 'reconstruction of the ZZ-Top beetle'	1
…	…	ZZF-Information 'ZZF information'[5]	1

Table 1: Lemma types and their frequencies from the *Stuttgarter-Zeitung* Corpus

frequency	frequency of frequencies	frequency	frequency of frequencies
1	404,579	…	…
2	96,981	708,531	1
3	43,357	710,719	1
4	26,159	1,605,763	1
5	17,559	1,848,517	1
…	…	3,571,573	1

Table 2: Frequency spectrum of *Stuttgarter-Zeitung* Corpus

When we consider that the *Stuttgarter-Zeitung* Corpus contains 714,972 types altogether we see that more than half of these occur only once. This is, again, typical for naturally produced texts of any length: the median of lemma frequencies is almost always 1.[6] We have undertaken similar statistics for a 200 m word corpus – the results are qualitatively the same. Why is this interesting? Because it shows us that adding more text also adds more new words. Even if we produced a lexicon that contained the 715,972 lemmas from the *Stuttgarter Zeitung*, there would probably be new words in the newspaper issue of the next day.

[4] This was noticed in the early 20th century by George Kingsley Zipf among others and led to the formulation of the so-called Zipf's law which states that the most frequent type occurs twice as often as the second frequent type and three times as often as the third frequent type etc. A short discussion of Zipf's law and similar formulas can be found in Baayen (2001) and in Manning & Schütze (1999).

[5] Where ZZF stands for *Zentralverband zoologischer Fachgeschäfte* 'central committee of zoological stores'.

[6] Such a distribution is called an LNRE distribution (for Large Number of Rare Events). See Baayen (2001) for statistical models and techniques in dealing with LNRE distributions.

Kinds of new words

Consider the hapax legomena (words occurring only once) in Table 1: some of them are names (*Zywietz, Zyzik*), some are regular words from a genre that is not typically covered in a newspaper (here biology: *zytos, zytotoxisch*) and some are compounds (*ZZ-Top-Käfer-Nachbau*).

To illustrate the same point, Amsler (1984) carried out a comparison between the vocabulary found in a college dictionary (Webster's 7th) and a text corpus (New York Times News Service) and noted that the overlap was only 23% of the total vocabulary in either source. Three quarters of the 41% which only occurred in the corpus could be accounted for in terms of inflection, hyphenation at the end of a line, proper nouns, and obvious misspellings. Assuming that inflection is accounted for by a rule system, hyphenation is covered in a trivial pre-processing step, and misspellings are treated separately, proper nouns constitute an important source of incompleteness. What happens with the remaining quarter? Amsler states that these cases cannot be classified without individual inspection.

From looking at Table 1 it can be predicted, however, that – in addition to names and misspellings – at least the following classes are represented in such a sample:

a) words missing in the lexicon because they belong to a genre that is not covered by it (obsolete language, scientific language, specialized language, …)
b) words belonging to a part of the text that is written in another language (quotations)[7]
c) neologisms introduced to name a new concept or to provide an alternative name for an existing concept. The latter may occur for all kinds of sociolinguistic reasons (jargon, sociolects, 'fashion', creativity in advertising, etc.)
d) …

These reasons have to be treated differently. Forms missing because of a) can, in principle, only be treated by enlarging the lexicon (this is at least true for simplex forms and lexicalized forms, see below). Forms missing because of b) can be omitted for many applications, if they need to be recognized they also have to be added to the lexicon, or another lexicon has to be added. This leaves neologisms: Here we have to distinguish between simplex elements that come through borrowing, creativity (e.g. German *Afronaut* for the first space tourist from Africa), clipping (e.g. German *kindi* from *Kindergarten*) etc.– again these can only be listed in the lexicon – and complex words. Complex words are the result of the operation of word formation rules. In the next section we give a very basic introduction to word formation and morphological productivity.

Basic word formation: elements and rules[8]

Morphology is traditionally divided into inflection and word formation. Intuitively, inflection is the formation of word forms of a lexeme for the appropriate syntactic

[7] This is quite common: think for example of English song texts, film titles, advertisements or quotations in German or French newspaper text.

[8] In this tutorial we do not have the space to explain word formation in a lot of detail. Therefore, many of the problematic (and interesting!) cases and issues cannot be touched upon. For more comprehensive introductions to word formation see Bauer (1988), Spencer (1991), Carstairs (1992). You can find introductions and overviews for a lot of issues in Spencer & Zwicky (1998) and Booij, Lehmann & Mugdan (2000). We also have to warn you that we need to gloss over a number of really problematic definitions.

environment (e.g. *pushes* for *push*, *craignent* for *craindre*) whereas word formation is the formation of new lexemes. Although this distinction is by no means uncontroversial and it is difficult to classify certain borderline cases (e.g. participles, the comparative and superlative of adjectives), we will assume here that it is possible to distinguish the two. An overview of the discussion is given in ten Hacken (1994), a brief summary from a somewhat different perspective is found in Booij (2000). Here we concentrate on word formation.

There are different ways of conceiving the rules of word formation. Consider the processes in (1) and (2), corresponding between English (a) and French (b).

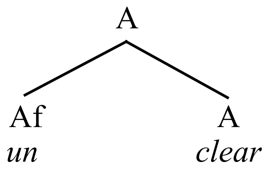
(1) a. clear unclear
b. fidèle infidèle

(2) a. sing song
b. fleur floral

Hockett (1954) outlines two models of describing what we can observe in these examples. In the Item and Arrangement (IA) model, (1a) is the combination of a stem *clear* and a prefix *un*. This combination results in the concatenation of the two forms and the compositional combination of syntactic and semantic properties. In (1a) the result is a form *unclear* with the syntactic category Adjective and the meaning of an antonym of *clear*. The examples in (2) require additional rules to modify the stem. In (2a), stem vowel change is the only thing marking the difference between the two words. The alternative model is called Item & Process (IP). In this model, word formation rules are processes applying to a base. In (1a), the process adds *un* to the left of the stem *clear*. In (2a), the process changes the stem vowel of *sing*. In IP there are no morphemes but only lexemes and processes. In modern morphological theories both are represented, e.g. Lieber (1992) for IA and Anderson (1992) for IP.

An important difference for our purposes is that in IA we have a tree structure whereas in IP we have a derivation history. A tree structure represents the relationship between morphemes, e.g. (3). A derivation history lists the different stages rules applying, e.g. (4). It should be noted that there are many variants of IA and IP.

(3)



(4) *clear* ⇒ Antonym formation ⇒ *unclear*

Tree structures such as (3) are very convenient for the representation of concatenative processes, i.e. prefixation, suffixation, and compounding. Non-concatenative (aspects of) word formation rules are more difficult in this perspective. Examples are vowel change as illustrated in (2) and other changes as in the nominalizations of *recevoir* to *réception*, *probable* to *probabilité*, *conquérir* to *conquête* etc. Another type of example is conversion, where a new homophonous word with a different syntactic category is created. This is frequent in English, where nouns can become verbs (e.g. *house*) or the reverse (e.g. *break*). Various solutions have been proposed in an IA perspective. In an IP perspective, these problems do not arise, because prefixation and suffixation are considered as special cases of a more general rule of the type "affect the form of the input in … way and the meaning in … way."

The reason we are interested in word formation rules is their productivity. Productivity is a difficult and controversial concept, cf. Bauer (2001). Basically, a productive word formation rule can be used to produce new lexical items. When a speaker has a productive word formation rule at her disposal, she can use a word not in her mental lexicon and be understood as far as other speakers have the same word formation rule available. The productivity of word formation makes it impossible to cover the entire lexicon in a finite list.

Productivity is not quite the same as regularity. To the extent a word formation rule is regular, we can predict properties of the output on the basis of properties of the input. Regularity can be seen as a cline with a gradual transition from fully regular to completely irregular. Some interesting points on this cline are:

- semi-regularity, in which the output can be related to the output, but not predicted by it, e.g. abbreviations and clippings;
- a higher degree of regularity in which the form and meaning of the output can be predicted from the input, but not the application of the WFR, e.g. *unclear*, but not **undeep*;
- full regularity, in which also the existence of the output can be predicted, e.g. *-ing*-forms of English verbs.

The relationship between productivity and regularity is discussed by Corbin (1987). In a computational context, the regularity of a word formation rule is what makes it possible to describe it in the form of a procedure which can be used to recognize new words.

The productivity of word formation rules is, of course, responsible for many of the unknown words found in unseen text, as described in the previous section. Recall the examples: quite a few of the hapax legomena were compounds (*ZZ-Top-Hit*), compounding is a very productive process in German (and English), therefore we expect many of the unknown words to be productively formed compounds.

There are two approaches to assessing the productivity of a word formation rule: the qualitative and the quantitative approach. For a full understanding, a combination of both perspectives is necessary. In the qualitative approach, all the restrictions and linguistic properties of the rule are described (see the next paragraph for some examples). The qualitative description does not exhaust our intuition of productivity, there is also a quantitative element involved: intuitively one could say that some word formation rules seem to produce new words more readily than others – an intuition which cannot be formalized. In quantitative studies this intuition is approximated by the question: how probable is it that we will see a new type (lexeme) produced by word formation process X after we have sampled a certain amount of text? Quantitative studies of the productivity of word formation processes are important for the design of word formation systems if the resources are limited and one has to concentrate on the most productive word formation processes (on the quantitative aspects of productivity see for example Baayen 1992, 2000; Baayen & Lieber 1991; Plag 1999; for a discussion of some corpus related problems in calculating productivity indices see Evert & Lüdeling 2001).

There are two descriptive aspects of word formation that we want to cover here in a little more detail because they are important for the description of the word formation systems DeKo and Word Manager below. We use mainly German examples here because both DeKo and Word Manager deal with German word formation. However, both issues are relevant for other languages as well. First we talk about possible restrictions on word formation rules and second we describe some of the stem changes that occur in word formation.

Word formation rules can be restricted on all linguistic levels (for examples you can refer to any good descriptive work on word formation in the language you are studying; in German descriptive overviews are given in Fleischer & Barz (1992) and in Deutsche Wortbildung 1-5; in addition there are numerous studies on individual affixes). The following constraints all refer to productive rules; each affix has also formed lexicalized words which have to be memorized.

- syntax: Most (if not all) affixes attach only to stems of a certain part-of-speech. The German adjective-forming suffix *–bar* attaches only to transitive verbs: *essen* 'to eat' – *essbar* 'edible'. It corresponds to English and French *–able*. The English noun forming suffix *–ness* attaches only to adjectives etc.
- semantics: Some affixes attach only to stems with a certain semantic feature. For example, German circumfix *Ge- -e* does not attach to stative verbs (there is no **Gewisse* from stative *wissen* 'to know'). Conceptual information can perhaps be viewed as a subtype of semantic information – it is a strong factor in constraining word formation – especially if one considers the semantic regularities. For example, when the German adjective forming suffix *–lich* attaches to nouns denoting professions or relations the resulting adjective means something like 'from, by the N' or 'like the N': *richterlich* can mean 'by a judge' as in *eine richterliche Anordnung* 'an order by a judge' or 'like a judge' as in *eine richterliche Frisur* 'a judgelike hairdo'. When *–lich* attaches to nouns denoting a time span it means something completely different, namely 'every N': *stündlich* means 'every hour', *wöchentlich* means 'every week' etc.
- morphology: The morphological structure of the stem itself can play a role in the restriction of a word formation rule. *Ge- -e*, for example, only attaches to unprefixed verbs: *Gekaufe* 'repeated buying' from *kaufen* 'to buy' but not **Geverkaufe* from *verkaufen* 'to sell'.
- phonology: Many affixes attach only to stems with certain phonological features. The German noun suffix *–ei*, for example, attaches only to nouns that end in a schwa-syllable, otherwise the allomorphs *–erei* or *–elei* are used which introduce a schwa.
- origin: Sometimes the choice between two affixes depends on the perceived origin of the base. Thus, in English, *–ity* is rather attached to words of Latin or French origin whereas *–ness* has a wider distribution, e.g. *curiosity* but *broadness*. An example from German is the adjective forming suffix *–abel* 'able' which combines only with stems of learned or neoclassical origin: *repräsentabel* 'representable' from *repräsentieren* 'represent' but not **darstellabel* from *darstellen* 'to show, represent'.

There are probably many more factors that constrain word formation rules. These can only be found by careful empirical studies. Below you see how such factors can be used to constrain rules in a word formation system.

The second empirical point we need to mention are stem changes in word formation. In German, Dutch and some other languages, stems sometimes look different when they appear as non-heads in complex words than when they appear in isolation.[9] This can be due to (a) umlauts, (b) linking elements and (c) elision. Contrary to a common assumption, stem changes do not only appear in compounding but also in derivation.

(a) the back vowels *a*, *u*, *o* and the diphthong *au*, can be fronted (umlauted) if they appear in the last syllable of a non-head stem in a complex word. Examples are *Frau* 'woman' which has an umlauted form in *Fräulein* 'Miss' or *Stunde* 'hour' which has an umlauted form in *stündlich* 'hourly'.

(b) linking elements are elements that appear at the end of the non-head in compounds and derivations (turn to any German grammar for afull list). Examples are <n> as in *Katzenfutter* 'cat's food' from *Katze* 'cat' or <s> as in *Mitgliedsbeitrag* 'membership fee' from *Mitglied* 'member'.

(c) schwa syllables can sometimes be deleted in word formation, as in *sprachlich* 'linguistic' or *Sprachkurs* 'language class' from *Sprache* 'language'.

Such changes in the form of the base in a word formation process are also widespread in other languages. We have seen examples from French and English in (2) above. What is important to note here is that these stem changes are not always fully regular or predictable. That is, even phonologically and semantically very similar words do not always have the same changed stems in word formation. *Bund* 'union' and *Grund* 'basis' which belong to the same inflectional class, for example, look different in compounds, compare *Grundgesetz* 'constitution' and *Bundesgesetz* 'federal law' (example from Lüdeling & Fitschen 2002). Or consider *Frau* which is *Fräu* in *Fräulein* but *Frau* in *fraulich* 'feminine' while the compound *Jungfrau* 'virgin' is umlauted when it attaches to *–lich*: *jungfräulich*. Some stems have different forms, for example *Schwein* 'pig' can be found as *Schweine* in *Schweinebraten* 'pork roast', as *Schweins* in *Schweinsauge* 'pig's eye'. This makes it very difficult to treat this phenomenon automatically. As you see below, DeKo and Word Manager approach stem changes differently.

Word formation systems

Compared to the treatment of morphology in linguistic theory, the treatment in computational linguistics is marked by two features:

- Word formation is usually not a separate issue. It is integrated with inflectional morphology or ignored altogether.
- Morphological components are generally based on written text. Even in systems with spoken input or output, the morphological component works on an orthographic transcription.

The marginal position of word formation is illustrated by the treatment in general surveys of computational linguistics. Surveys such as Karlsson & Karttunen (1997) and

[9] Stem changes have been analysed in a number of different ways in the literature. Since many forms look like inflected forms (the so-called paradigmic forms) it is sometimes argued that these are inflected forms in word formation. There are good arguments against this view: (a) there are many stems that do not change in word formation, (b) changed stems do not necessarily have the semantics of the corresponding plural form and unchanged stems do not necessarily have the semantics of the singular, and (c) there are also non-paradigmic forms. See the discussions in Fuhrhop (1998) and Eisenberg (1998). The existence of stem changes is often used as an argument against the IA model, e.g. by Anderson (1992). An analysis in terms of “stem formation” is elaborated by ten Hacken (1994).

Sproat (2000a) do not even mention inflection and word formation as terms, let alone make the conceptual distinction. The starting point of the approaches they describe is clearly inflection. As far as word formation phenomena are treated, e.g. Sproat (2000a:50), they are not considered from the perspective of the creation of new lexemes, but as examples of more difficult combinations of formatives.

From a technical point of view, the domain of inflectional morphology is a rather well-explored area, in which most efforts are devoted to development rather than research. Techniques used are based on finite-state transducers as used originally in two-level morphology, cf. Sproat (1992) for an overview. Research concentrates to a large extent on complicated phenomena such as Arabic nonlinear morphology.[10]

Transferring the finite-state approach from inflection to word formation does not by itself cause many additional problems, but it does exacerbate a number of well-known problems of finite-state mechanisms:

- A finite-state rule system concatenates formatives from left to right.[11] As long as we are dealing with suffixation, there is no problem. For languages such as Finnish, Turkish, and Basque, which have only suffixation and a lot of it, finite-state morphology is ideal. Word formation in Indoeuropean languages, however, involves regular prefixation processes (cf. (1) above) in combination with more frequent suffixation.
- A finite-state rule system cannot express long-distance dependencies. A word such as *unacceptable* is problematic, because the prefix *un-* requires an adjective, but the adjectival status of *acceptable* depends on the suffix *-able*.[12] Therefore it is impossible to know whether *un-* can be prefixed until we arrive at the suffix.
- An important difference between inflection and derivation is the handling of information. In plural formation of nouns, e.g. *readers*, the feature plural is added to the lexeme *reader*. In the formation of *reader* from the verb *read*, however, the feature noun replaces the feature verb in the slot for syntactic category. Whereas *readers* is a plural noun, *reader* is not a noun-verb, but simply a noun.

The last problem is not necessarily linked to finite-state approaches, but in two-level morphology, cf. Koskenniemi (1983), Antworth (1990), the way features are handled is specifically geared towards inflection. The first two problems are inherent in finite-state technology and can only be handled by one of the following strategies:

- Giving up the finite-state constraint and using a more powerful rule type such as context-free rewrite rules. This implies a loss of computational efficiency, because context-free rules do not work in linear time.

[10] An example of this type of work is Kiraz (2001). In Arabic a root such as *ktb* is combined with a vowel pattern to produce words such as *kitaab* ('book') and *kutub* ('books'). It is interesting to note that the traditional approach to Arabic roots results in approximately 10,000 different items. This number corresponds more closely to the number of simple lexemes to be expected in the lexicon of a language than to the number of lexemes. It is then not surprising to find items such as *kaatib* ('writer'), *kutib* ('be written') with the same root.

[11] In principle we could of course reverse the entire system. Thus, languages such as Navajo, which use only prefixation, are not a major problem.

[12] There is of course a different prefixation process attaching *un-* to a verb as in *undo*, but it would yield the wrong analysis for *unacceptable*. The word means 'which cannot be accepted', not 'which can be unaccepted'.

- Using finite-state rules but hiding them in the linguistic specification interface. This implies compilation of a more powerful formalism into finite-state rules which do not express the rationality of the analysis.

Apart from such considerations of formalism, a more general strategy about the functionality of the word formation module has to be chosen. In this point there is a trade-off between precision of the analyses and investment of work in the specification. Two extreme positions can be characterized as follows:

- Word formation is treated as a rule component only. The lexicon contains only simple entries.
- Word formation is treated as a property of lexicalized items. The lexicon contains the analysis in terms of word formation rules for each complex lexeme.

The first strategy is problematic in the sense that word formation rules create a lot of ambiguity which can only be resolved by considering which of the possible analysis has been lexicalized. The second strategy is deficient if it does not include a rule component which can deal with non-lexicalized words. As shown above, the processing of unseen text requires this possibility and this is one of the main reasons to have a word formation component in the first place.

In practice, the strategy adopted by a word formation system can be placed somewhere on the cline between these two positions in view of the number of lexical items included and the amount of information restricting the application of word formation rules. This will be illustrated in the discussion of two systems, DeKo and Word Manager.

DeKo

DeKo (for **De**rivation and **Ko**mposition, funded by the state of Baden-Württemberg from Jan 2000 – June 2001) is designed as a German word formation component in a larger computational linguistic application.[13] It was done on a much smaller scale than Word Manager and is not used in any commercial products. In this tutorial we focus on some basic design features – especially where they differ from Word Manager's features. More details can be found in Heid 2001, Schmid et al. 2001 and at http://www.ims.uni-stuttgart.de/projekte/DeKo/. Here we want to concentrate on the corpus-based acquisition of data, the item-and-arrangement design, analysis and structure, and the interaction between DeKo and the lexicon. Although the DeKo project proper is finished, work is still being done to improve the program and especially to extend the lexicon.

Acquisition of data

As indicated above, word formation can be restricted on all linguistic levels. DeKo is designed so that it can, in principle, use all this information in its rules in order to minimize ambiguities (remember that DeKo is used as a component in applications that deal with unseen text and that the analyses can therefore not be manually corrected). Although there is a lot of descriptive literature on German word formation, there is no

[13] At the moment it is used in the Text-To-Speech system IMS-FESTIVAL within the SmartKom Project and in a terminology extraction project. There are plans to use it as a backup-program for unknown words with the German LFG parser in the ParGram project.

standardized collection of the relevant data that we could use. Therefore we first collected and systematized the data ourselves (using every available source, of course).

Data acquisition in DeKo was done on the basis of a corpus: we used German newspaper corpora, which were tagged with the TreeTagger (Schmid 1994) and lemmatized with DMOR (Schiller 1996), for searching and pre-processing we used the Corpus Query Processor (Schiller 1996) and a number of Perl scripts. In acquiring and systematizing the data we made a distinction between word formation involving selecting elements (roughly derivation) and word formation involving only categories (compounding). For expository purposes we concentrate on a derivation process here and only briefly describe a compounding process below.

We described each derivational process on three levels. First we collected information about the selecting element (affix) itself. Table 3 shows the entry for the diminutive suffix *–chen.* Comparable information is collected for each affix.

In a second table we collect information on the different productive word formation patterns for each affix, since they can have different properties. *–chen*, for example, attaches productively to nouns, names, and adjectives. The adjective-forming suffix –*lich* attaches to nouns, adjectives, and verbs: the restrictions are quite different for each pattern. The Ns that attach to *–lich*, for example, may be morphologically complex (e.g. *Amtsarzt* 'public health officer' in *amtsärztlich*) but the adjectives in the Adj+*–lich* pattern can only be morphologically simplex. For each pattern, we distinguish between information on the bases that the affix attaches to and information on the whole complex word.

The information that has to be collected for composition patterns looks a little different because here we do not have a 'selecting element'. That means that we can formulate the restrictions only with respect to the categories that are involved. Noun+noun compounding is productive and recursive and relatively unrestricted but other compounding patterns are highly restricted. Adjective+adjective compounding, for example, typically do not have a complex head, and noun+verb compounding is non-recursive. Information about compounding patterns is collected in tables that are similar to the derivation tables.

Examples for such tables can be found at
http://www.ims.uni-stuttgart.de/projekte/DeKo/bspdaten.shtml.
At the moment DeKo has collected data for about 300 selecting elements (affixes) and 20 composition patterns.

	produces	token/ types/ hapaxes[14] (in 200m words)	origin	semantic function	umlauted bases?	classical bases?	other foreign bases?	Bound bases?[15]	abbreviations as bases?	names as bases?	phrasal bases?	morphologically complex bases?
chen SAMPA: C@n	N	94825 1199 394	native	diminutive (nouns, adjectives) hypocoristic (names)	y	y	y	N	y	y	n	y, compounds and derivations
examples for illustration					Häuschen 'little house', Blümchen 'little flower'	Persönchen 'little person', Reförmchen 'little reform' Progrämmchen 'little program'	Computerchen 'little computer', Affairchen 'small affair'		BITchen, DMchen	Kläuschen, Mariechen		Fangfrägelchen 'little trick question', Muttersöhnchen 'namby-pamby boy'

…	bears stress?	influences stress?	prohibits final devoicing?	causes resyllabification or other phonological changes	influences argument structure of base?[16]	corresponding affixes
chen, continued	n	n	n	elision of final stem syllable possible (schwa syllable)	n	-lein
examples				Äffchen 'small monkey' from Affe Brünnchen 'small fountain' from Brunnen		

Table 3: the diminutive suffix *–chen*

[14] Hapax legomena are words that occur only once. They are often used as an indication for the productivity of a word formation process.

[15] Bound bases are found in neoclassical compounds. Examples are *anthropo*- and *–phagie* in *Anthropophagie* 'cannibalism'. These elements behave semantically like stems but do not occur free.

[16] This is applicable mainly to verbal prefixes which can influence the argument structure of their base verb. An example is the prefix *be*- which can transitivize intransitive verbs, compare intransitive *trauern* 'to mourn' to transitive *betrauern* 'to mourn sb'.

In addition to the productive and regular cases, we collect lists of 'lexicalized' or 'semi-regular' words, that is words that are either not fully regular and have to be memorized or words that belong to a regular unproductive pattern and can be listed as such. An example for the first kind of lexicalized words would be *wöchentlich* 'weekly' which is derived from *Woche* 'week' and fits perfectly into the semantic pattern of N denoting time span + *–lich* but is phonologically irregular in that it contains a <t>. An example for the second kind is the list of verbs of killing and dying beginning with the prefix *er-*: *erschlagen* 'to strike dead', *erwürgen* 'to strangle, *erstechen* 'to stab' etc. This list has about 20 entries which are somehow semantically related but it is not possible to form any new words of killing using this pattern. Therefore we don't want to write a productive rule for these words. Semi-regular words receive morphological analysis and structure.

Finally we collect a list of words that look like they could be complex but are really simplex. These have to be listed as simplex words in the lexicon, otherwise the analyser would give them structure. Examples are *wichtig* 'important' which is not *Wicht* 'gnome' + *–ig* or *freilich* 'really' which should not be analysed as *frei* 'free' + *–lich*.

Word formation stem forms

DeKo uses a strict Item and Arrangement (IA) approach for derivation and compounding: morphological elements are concatenated according to the word formation rules.

As discussed above, there are stem changes (umlaut, linking, elision) in German word formation. These make the analysis of complex words more difficult. The <n> in *Anzeigenadel* could, for example, belong to *Nadel* 'pin' or it could be a linking element for *Anzeige* 'advertisement, display'; the compound could then either be *Anzeige+Nadel* 'display needle' (as in a speedometer or a scale) or *Anzeigen+Adel* 'advertisement nobility'. The latter reading is improbable but word formation systems do not normally contain a semantic component. And it is, of course, not an illegal or impossible reading.

In an Item and Process approach stem changes are dealt with via rules. Since stem changes in German are not regular or predictable we follow Fuhrhop (1998) and Eisenberg (1998) on listing the forms that an element may take. These forms are called word formation stem forms. Free and bound morphological elements have one or more derivation stem form(s) and one or more compounding stem form(s). Very often these stem forms look just like the regular stem. Consider the examples in Table 4 where we have listed the word formation stem forms of some morphological elements together with the appropriate examples:

stem	Frau 'woman'	-keit (noun forming suffix)	***les- 'to read'***
derivation stem forms	fräu- frau-	-	les
derivation examples	Fräulein 'Miss' fraulich 'feminine'	-	lesbar
compounding stem forms	frauen-	keits	lese
compounding examples	Frauenzeitung 'women's newspaper'	Sauberkeitsfimmel 'cleanliness mania'	Leselampe 'reading lamp'

Table 4: word formation stem forms for some morphological elements

Here you can see that the suffix *–keit* has no derivation stem form. That means that it cannot be used in further derivation (it is a so-called closing suffix, see Aronoff & Fuhrhop 2001). It has a compounding stem form, however. The word formation stem forms are propagated in complex words using these elements: each complex word ending in *–keit*, for example, will automatically have a compounding stem form ending in *–keits*.[17]

Such an approach minimized the ambiguities that stem changes can cause. However, one has to acquire all the word formation stem forms. For our lexicon this is done semi-automatically as described in Heid, Säuberlich & Fitschen (2002).

Analysis & structure

The word formation rules build on the information collected in the tables and lists as well as on the information stored in the lexicon (see next paragraph). The rules first make a morpheme analysis. In addition to a morpheme analysis the DeKo project wanted to provide a hierarchical structure to the complex words (none of the other word formation projects for German that we are aware of does this). Hierarchical structure can add a lot of information to the morpheme analysis – it is especially important because it features in pronunciation rules (TTS system), but it is also useful in information retrieval.

Interaction between DeKo and the lexicon

The DeKo rules can only work if they can refer to detailed information on lexical items – therefore the DeKo team and other researchers at the IMS in Stuttgart developed a highly flexible lexicon concept where different kinds of information are stored together with morphological elements (see Lüdeling & Fitschen 2002 for more details).
At the moment the relevant information is still being collected and encoded into the IMSLex. Therefore, the DeKo rules as they stand now are much less specific than they should be.

The following analysis for the noun *Abörtchen* 'little toilet' which should have been analysed as *Abort*+*-chen* can be prevented if the noun prefix *a-* has the restriction that it only combines with neoclassical elements and if all elements in the lexicon are marked by origin.

```
Abörtchen
{{A[npref][++]Borte[nomen][sg][fem][stem]}[pref_derivat][++]chen[suff][
nomen][pl][neut][stem]}[suff_derivat]
```

Very often we find many legally possible analyses, where one is the most plausible one (this is the same situation that we find in many syntactic parsers). Each word beginning with a verb and ending in <bar>, for example, is ambiguous between at least three analyses: a derivation with the adjective suffix *–bar*, a compound with the noun *Bar* 'bar' and a compound with the adjective *bar* 'in cash'. The compound analyses are possible because there are general rules that allow V+N compounding (as in *Denkrichtung* 'direction of thought') and V+Adj compounding (as in *denkfaul* 'too lazy to think') Typically the derivation is the most plausible reading and the adjective compound reading is almost never correct. But sometimes the noun compound reading might be the intended reading. What do we do when such a situation arises?

[17] Exceptions to this rule must be listed.

There are three principal strategies: one could manually choose the plausible reading and code it in the lexicon, one could leave the decision to an automatic choice function (for example one that associates 'costs' or probabilities with rules)[18] or one could accept all legal analyses and let another component decide. We have chosen to combine the first and the last strategy, so that the application can either rely on manually corrected lexicon information alone or it can accept all legal analyses.

Implementation

DeKo is implemented as a series of finite-state transducers, using the FST-suite provided by AT&T (Sproat 2000b). A more detailed description of the architecture is given in Schmid et al. (2001) and examples for the rule format are provided in Säuberlich (2001). We need to model three types of rules:

- The sequential analysis into morphemes is done in a declarative grammar. For example, the adjective *unregierbar* 'ungovernable' which has to be divided into the morphological elements *un-* + *regier* + *-bar* can be treated by the following grammar:

  START PREF un[adj.pref] +
  PREF STEM regier[verb]
  STEM SUFF + bar[suff][adj]

  where the parts in the square brackets provide the restrictions. Here we can formulate restrictions on all relevant linguistic levels as long as the information is present in the lexicon.
- The hierarchical structure is provided by a context-free grammar which adds the appropriate brackets to the morpheme analysis. As described above, a declarative grammar cannot describe long-distance dependencies, such as the fact that the prefix *un-* wants to combine only with adjectives. Therefore the first step would give us a number of incorrect analyses. These can be thrown out at this step.
- Finally, phonological mechanisms are coded as context-sensitive rewrite rules. Since we use a strict IA approach with word formation stem forms, we do not have to code rules that eliminate linkers, change umlauts etc. However, as long as the lexicon is not fully acquired we have these rules as a fall-back strategy.

Word Manager

Goals

Word Manager (WM) is a system for morphological dictionaries. It is not primarily a system for the treatment of word formation, because its scope is much broader. Its domain is defined so as to encompass word formation, however, so that in view of the relative rarity of word formation systems it is worth considering WM from this perspective.

[18] It is, of course, difficult to train or find the appropriate choice function. Since the same issue is problematic in parsers the strategies of dealing with too much ambiguity can be transferred from there.

The original motivation for WM stems from the discovery of the so-called "lexical bottleneck". In the course of the 1980s it was realized that the quality of many of the intricate rule systems developed in computational linguistics could not be used in practice because they did not have a lexicon, cf. Ritchie (1987). The obvious solution was of course making lexical resources reusable.

The originality of WM lies in its approach to reusability. Ten Hacken & Domenig (1996) present this approach in terms of the diagram in Fig. 1.

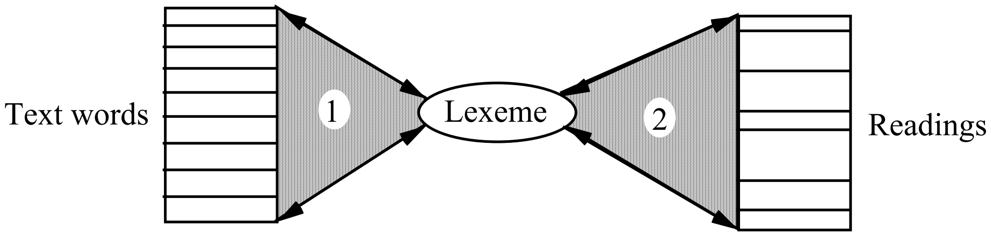


Fig. 1: The bow-tie model adopted by WM.

The central position in the bow-tie model is taken by the lexeme. The notion of lexeme in WM is similar to the one adopted by Matthews (1974), Aronoff (1994) and others. A lexeme is a word considered as an inflectional paradigm. Fig. 1 highlights two mappings involving the lexeme:

1. the mapping between the lexeme and a list of text words, i.e. forms as they appear between spaces and punctuation marks in a text;
2. the mapping between the lexeme and a list of readings, i.e. words with syntactic and semantic analysis as required by the theory and application of a system of computational linguistics.

The independence of the two mappings is illustrated by classical cases of homonymy, e.g. *bank*. Depending on the type of application, different readings will be required for the financial institution, the building it is housed in, an elevated section of the seabed, the border of a river, etc. In all these cases, there is a singular form *bank* and a plural form *banks*. This means that in terms of Fig. 1, *bank* as a noun is a lexeme mapped to the singular and plural forms in mapping 1 and the different readings in mapping 2.

The goal of WM is to provide a reusable resource for mapping 1. This mapping is taken to comprise the following cases:

1. The text word is a wordform in the paradigm of a lexeme in the database.
2. The text word is a wordform in the paradigm of a new lexeme, resulting from the productive application of word formation rules.
3. The text word has to be split up before mapping to a lexeme is possible.
4. The text word has to be combined with other text words before mapping to a lexeme is possible.

The simplest type is case 1, where a text word can be attributed to a lexeme. Of course there are ambiguities here, e.g. *nuit* as a noun ('night') or as the third person singular of *nuire* ('damage'), but the operation is a matter of classification only. Case 2 concerns unseen words analysed by word formation. Cases 3 and 4 involve operations on the string of text words preceding the classification process. An example of case 3 is *a-t-il* analysed as the third person singular of *avoir* followed by the personal pronoun *il*. Case

4 applies to multi-word units in the orthographic sense. This includes a considerable part of word formation, for instance most English compounds, cf. *fire brigade*.

Ten Hacken (1999) describes the opposition between the coverage of the lexical component in the standard approach and in WM in terms of two orthogonal dichotomies:

- The standard approach distinguishes the lexicon from the grammar, such that information about individual entries is in the lexicon whereas rules are in the grammar.
- The WM approach distinguishes between the two mappings in the bow-tie model, such that information for the mapping between text words and lexemes (both rules and entries) is in the WM database, whereas information for the mapping between lexemes and readings should be covered in a different component.

The two perspectives can be represented as in Fig. 2:

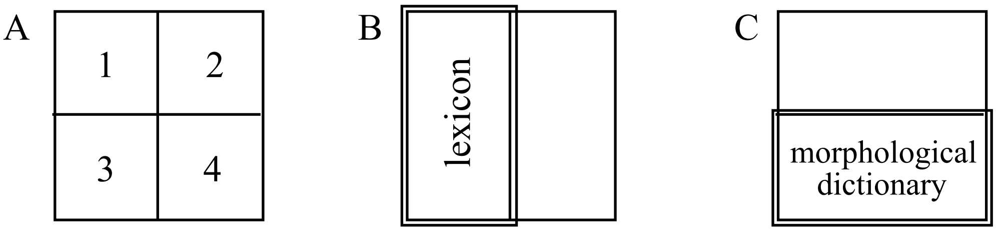


Fig. 2: Coverage of WM resources (C) as opposed to more traditional lexicons (B).

Somewhat simplified, the areas in Fig. 2A can be characterized as follows:

1. syntactic and semantic lexical knowledge,
2. syntactic and semantic rule knowledge,
3. morphological lexical knowledge, and
4. morphological rule knowledge.

The combination of lexical knowledge and rule knowledge enables WM to function as a full morphological component. As shown by ten Hacken (1998), the effects of this coverage are particularly striking in the domain of word formation, for which a system taking a lexicon as modelled in Fig. 2B as a basis lacks the procedural component. Thus a formalism such as DATR, as described by Evans & Gazdar (1996), though able to represent word formation relationships, cannot deal with unseen words without a separate recognition module. In WM, word formation rules are at the same time available declaratively, as the structural backbone of the database, and procedurally for the recognition of new words.

History

WM is a long-term, open-ended project which originated with Domenig (1989). Subsequently it was developed at Universities in Basel, Amsterdam (Vrije Universiteit), and Lugano (SUPSI and USI), funded in part by the Swiss federal government and by private companies.

In order to describe the history of the development of WM, the best starting point is the development stages leading to a directly applicable component which can be integrated in a real-life environment. There are three main stages involved:

1. The morphological system of a language (inflection and word formation) is described in terms of the WM formalism. The result of this linguistic specification phase is a (morphological) rule database.
2. The lexemes of a language are described in terms of the rules in the rule database. The result of this lexicographic specification phase is a (morphological) lexicon database.
3. A dedicated finite-state tool is developed on the basis of the lexicon database. The result is a small, efficient module performing a specific task independently of the lexicon database.

At the start of the project, the computational environment was developed. First the formalism and compiler for the mappings in which text words need not be split up or combined was implemented and tested. The formalism is described in Domenig & ten Hacken (1992). This core WM system served as a basis for a number of Ph.D. dissertations in which further components were added. The extension for the treatment of clitics and multi-word units, Phrase Manager (PM), is described in Pedrazzini (1994). A module for the use of word formation rules as a basis for the semi-automatic classification of new entries is described by Hsiung (1995). This module proposes a number of analyses for a new word on the basis of word formation rules and existing lexemes and formatives. Holz (1995) developed an interface for the specification of entries. For other parts, documentation was not extensively published. The mechanism for deriving finite-state tools from lexicon databases is described by Pedrazzini & ten Hacken (1998) and Pedrazzini (1999).

WM was conceived as a language-independent system. Although the formalism is very flexible, the system is not equally adapted to all languages. It works best with languages having a morphological system such as Germanic and Romance languages, with a moderate amount of inflection and non-concatenative processes. The first morphological rule database to be developed was the Italian database described by Bopp (1993). Other complete rule databases were developed for German and English. The German rule database uses PM for the analysis of separable verbs, such as *aufhören* in (5):

(5) a. Anna glaubt, dass Bernard aufhört.
(‘Anna believes that Bernard stops’)
b. Claudia hört jetzt auf.
(‘Claudia stops now PRT’)
c. Daniel versucht aufzuhören.
(‘Daniel tries to_stop’)

As described by ten Hacken & Bopp (1998), the interaction of inflection rules, word formation rules, clitic rules (for (5c)), and rules for multi-word units (for (5b)) makes it possible to analyse all occurrences of *aufhören* in (5) as instances of the same lexeme. For English, Tschichold (2000) describes the rules for the analysis of multi-word units more generally.

The first lexicon database to be developed was for German. It has now reached a size of 200,000 lexemes. The development of English and Italian lexicon databases was undertaken in a parallel effort, funded by the Swiss National Science Foundation. The parallelism entails that the same lexicographic guidelines are used for both languages. Word formation plays a central role in the development, as discussed by ten Hacken (2002) and ten Hacken & Smyk (2002).

Linguistic Approach

The basic entity in a WM database is the formative. A formative is a combination of a string and a set of features. There are three types of rule for manipulating formatives: IRules, WFRules, and SRules.[19] IRules and WFRules combine formatives in slightly different ways. SRules change the form of formatives.

In an IRule, a set of stems and a set of affixes are combined to produce the inflectional paradigms for the lexemes with the stems as their base. The combination of formatives in an IRule leads to the concatenation of the forms and the combination of the feature sets. An example from the Italian rule database is (6).

```
(6)  (ICat V-base.pres.1st) (ICat V-suffix.1st)(Mod Ind)(Temp Pres)
```

The statement in (6) is part of the rule for first conjugation verbs, e.g. *cantare* ('sing'). It consists of two parts separated by a tab. The first part characterizes a class of stems, first declension verbs, and the second part a class of endings, the first declension present indicative endings. Thus, (6) is responsible for the paradigm *canto, canti, canta, cantiamo, cantate, cantano* ('I/you/she/we/you/they sing').

In a WFRule, there are no paradigms, but only individual applications of a word formation process. Moreover, there is more freedom in the specification of the result of rule application. An example, again taken from the Italian rule database, is given in (7).

```
(7)  source
     (Cat Adj)(Manner Qual) (?IRule ?)
        1    (ICat A-Base)
     2  (WFCat Suffix)

     target
     (RIRule 1st-Conjug)
        1 2  (ICat V-Base)
```

The WFRule in (7) produce the verb *biancheggiare* ('whiten') from the adjective *bianco* and the suffix *-eggi*. In the source specification, two numbered items are identified, the first a qualitative adjective and the second a suffix.[20] In the target the result is specified as a first declension verb produced from the concatenation of the forms of the two source elements. It is also possible to propagate information from the source to the

[19] In addition there are special rule types for clitics (CRules) and multi-word units (four types with different functions). They are important for the treatment of certain word formation processes, e.g. German separable verbs as illustrated in (5) and English multi-word compounds such as *fire brigade*. The latter type does not add anything conceptually relevant apart from the fact that it is possible to treat *fire brigade* as a morphological formation on a par with *girlfriend*. Separable verbs constitute a rather complex phenomenon in a text-based system because of the interplay between one-word and multi-word forms of the same lexeme. The treatment of the phenomenon is explained in detail in ten Hacken & Bopp (1998).

[20] The selection of the suffix is restricted by the position of the rule in the word formation tree. For details cf. Domenig & ten Hacken (1992).

target by marking the relevant feature in the source. Thus for the verbal prefix *ri-*, corresponding to English *re-*, the inflection class of the base is passed on to the target. In this way the verb *riandare* ('go again') is inflected in the same, highly irregular way as *andare* ('go').

The differences between IRules and WFRules express the differences between inflection and word formation as they are analysed in WM. Whereas inflection is the paradigmatic realization of inflectional features for a lexeme, word formation is the creation of a new lexeme according to the constraints specified in the WFRule. WFRules work on a case by case basis. The attribution of a stem to an IRule entails the existence of the entire paradigm. The application of a WFRule creates a single new lexeme.

SRules are the only rules which may change the form of a formative. General processes, such as the addition of an *h* to preserve the pronunciation in *biancheggiare*, are treated at a central level for the entire language. Specific cases, such as the choice in German between umlauted or non-umlauted results in suffixation with *-lich* (cf. *fraulich* and *jungfräulich* above) can be treated by having different WFRules, one of which invokes an umlaut SRule.

The choice of concatenation as the basic operation in rules (6) and (7) may suggest that WM adopts an Item & Arrangement approach to morphology. This impression is not quite correct, however, because stems and affixes are treated as different types of entity. Formatives are divided into fully specified formatives and underspecified formatives. The names are chosen from the perspective of the rule database. Fully specified formatives, such as affixes, are specified in the rule database and cannot be changed in the specification of the lexicon database. Underspecified formatives are specified only in terms of their class properties, e.g. (ICat A-Base), in the rule database. Lexicographic specification instantiates these classes by adding a form and (when applicable) idiosyncratic properties.

The lexicographer applying the WFRule in (7) selects an adjective stem and one of the suffixes *-eggi*, *-ific*, *-it*, or *-izz*. Therefore the rule database presents itself as a set of processes and WM can arguably be thought of as adopting an Item & Process model of morphology. This impression is reinforced by the automatic application of SRules. In cases such as *biancheggiare* and *riandare*, the lexicographer does not have to specify anything to trigger the SRules adding the *h* in the former and producing the irregular inflectional forms of the latter. In the case of German *-lich*, the choice between two WFRules automatically determines which suffixation process is applied, the one with or the one without umlaut.

Practical Use

For a system aiming to make available a set of reusable dictionaries, it is of course essential to specify in what sense and under what conditions these dictionaries can be used. As mentioned above, there are three stages in the development of a WM lexical component. The difference between a rule database (with only a few sample entries) and a lexicon database (with a large set of entries) is clear. Why do we need a further step after the development of the lexicon database? The difference between a WM lexicon database and a lexical tool based on it can be characterized as follows:

- The lexicon database requires the entire WM application for its running. A lexical tool runs independently.

- The lexicon database is a large file running only on a MacIntosh computer with large amounts of RAM (1 GB recommended). A lexical tool is a fast-running, small (less than 2 MB) finite-state transducer. Tools are platform-independent in the sense that a tool with a particular functionality can be derived from the database for any platform desired.
- The lexicon database contains all the information about inflection and word formation rules and relationships for a language. It is an object-oriented database with high flexibility. A lexical tool is dedicated to a particular task, using a selection of the information in the database in a particular way. Tools are very specific, for instance in the format of their output.

An example of the implications of these differences can be taken from the use of word formation in WM in the context of Computer-Assisted Language Learning. Ten Hacken (1998) compares the treatment of word formation in WM with its treatment in learner's dictionaries published as books and in a number of other electronic dictionaries which do not have a word formation component. The conclusion is that the WM treatment offers possibilities unknown in the other contexts. Ten Hacken & Tschichold (2001) describe these possibilities more concretely, concentrating on the information available in the WM lexicon database and the browsers giving access to this information. Thus, it is immediately visible how many lexemes in the 200,000 entry database for German were formed by a particular process or set of processes, or which entries are derived from the noun *kind* ('child'). The result of developing a number of dedicated lexical tools for CALL can be seen at `http://www.canoo.net/`. Here it is also possible, for instance, to have one's text checked for one of the recognized style sheets implementing the German spelling reform.

Another domain in which WM databases have been used is terminology. In collaboration with the UBS bank, a module was developed which recognizes banking terminology in unseen text and provides an on-line link to the relevant entry in a terminological database, cf. Zappatore & ten Hacken (2000). Here the WFRules are used not only as a structuring device of the terminological lexicon, but also as a way for recognizing terms when they are 'hidden' in nominalizations, compounds, etc. Thus, *Verwaltungsrat* ('board of directors') is also recognized in *Verwaltungsratsvakanz* ('vacancy in the board of directors'). One of the reasons why WM is particularly suited to this task in a multilingual system (German, English, Italian) is that it can treat single-word and multi-word terms equally.

As a final example, Pedrazzini & ten Hacken (1998) describe the prototype of a "generative spellchecker". In particular in German, where compounds are written as one word, spellcheckers regularly fail on words such as *Nordostgrenze* ('north-eastern border') because they are not in the dictionary. The generative spellchecker recognizes these words as possible words. Depending on the elaboration of the prototype they may be proposed as a list to be checked for spelling errors[21] or for enlarging the lexicon database. Applied to a German corpus of newspaper text with a 100,000 entry lexicon, 7% of the words in the corpus were recognized with the generative component of the spellchecker.

[21] Some possible words are more likely to be spelling errors than intentional coinings. Thus, *Leimwand* is a possible compound of *Leim* ('glue') and *Wand* ('wall'), but more probably a spelling error for *Leinwand* ('film screen').

Pius ten Hacken, WWZ – Abt. Geisteswiss. Informatik, Universität Basel,
Petersgraben 51, CH-4051 Basel, Switzerland, pius.tenhacken@unibas.ch

Anke Lüdeling, Inst. für Kognitionswissenschaft, Universität Osnabrück,
Katharinenstrasse 24, D-49069 Osnabrück, Germany, aluedeli@uos.de

Appendix: Word formation Systems

Here we have collected references and urls of some word formation systems, ordered by language:

English

- ALE-RA http://nl.ijs.si/et/Thesis/ALE-RA/
- The Unified Medical Language System UMLS provides a morphological variation system for English medical terms http://www.nlm.nih.gov/research/umls/

German

- DeKo (for Derivation und Komposition, IMS, University of Stuttgart) http://www.ims.uni-stuttgart.de/projekte/DeKo
- Projekt Deutscher Wortschatz (University of Leipzig): http://wortschatz.uni-leipzig.de
- Deutsche Malaga Morphologie (University of Erlangen): http://www.linguistik.uni-erlangen.de/~orlorenz/DMM/DMM.html
- CISLEX (University of Munich): http://www.cis.uni-muenchen.de/projects/CISLEX:html
- GerTWOL (Lingsoft Inc.): http://www.lingsoft.fi/cgi-bin/gertwol
- and there is a German version of WordManager (University of Basel & Canoo) http://www.wordmanager.com

Italian

- IMMORTALE (University of Venice), Information and publications can be found at http://project.cgm.unive.it

Norwegian

- Oslo-Bergen-taggeren http://decentius.hit.uib.no:8005/cl/cgp/test.html

Romanian

- PARADIGM
  Tufis D., Popescu O., "A Unified Management and Processing of Word-Forms, Idioms and Analytical Compounds", in Jurgen Kunze and Dorothy Reinman (eds.), Proceedings of the 5th EACL, Berlin, 1991, pp.95-100
- the MICH classification-based system
  Dan Cristea (1994): The Classification Language MICH, Research Report, LIMSI-CNRS, Universite Paris-Sud, Orsay.

Russian

- RUSLO (abbreviated from the Russian "RUSskoye SLOvoobrazovaniye" = "Russian Derivation") http://194.226.57.46/uvk1838/Sciper/volume1/pertsova.htm

Turkish

- http://www.sabanciuniv.edu/fens/people/oflazer/

Multilingual

- Word-Manager (German, English, Italian, ...)[22] http://www.unibas.ch/LIlab/projects/wordmanager/wordmanager.html
- Lingsoft (a Finnish company) provides morphology systems based on two-level morphology for the following languages: Finnish, Swedish, Norwegian (bokmål), Danish and German. Although it treats mainly inflection, there are some derivational components as well. www.lingsoft.fi

More general information about morphology systems (dealing mostly with inflection) can be found

- http://www.sil.org/computing/comp-morph-phon.html
- http://www.xrce.xerox.com/competencies/content-analysis/fsnlp/morph.en.html

References

Amsler, Robert A. (1984), 'Machine-Readable Dictionaries', *Annual Review of Information Science and Technology* 19:161-209.

Anderson, Stephen R. (1992), *A-Morphous Morphology*, Cambridge: Cambridge University Press.

Antworth, Evan L. (1990), *PC-KIMMO: A Two-level Processor for Morphological Analysis*, Dallas (Texas): Summer Institute of Linguistics.

Aronoff, Mark & Fuhrhop, Nanna (2001), 'Restricting suffix combinations in German and English: closing suffixes and the monosuffix constraint', *Natural Language and Linguistic Theory*.

Aronoff, Mark H. (1994), *Morphology by Itself: Stems and Inflectional Classes*, Cambridge (Mass.): MIT Press.

Baayen, Harald & Lieber, Rochelle (1991), 'Productivity and English derivation: a corpus-based study', *Linguistics* 29:801-843.

Baayen, Harald (1992), 'Quantitative aspects of morphological productivity', in *Yearbook of Morphology* 1991:109-149.

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

Bauer, Laurie (1988), *Introducing Linguistic Morphology*, Edinburgh: Edinburgh University Press.

[22] Note that the server is case-sensitive and that the first two characters of "LIlab" are capitals, the rest of the url lower case.

Bauer, Laurie (2001), *Morphological Productivity*, Cambridge: Cambridge University Press.

Booij, Geert (2000), 'Inflection and Derivation', in Booij et al. (eds.), 360-369.

Booij, Geert; Lehmann, Christian & Mugdan, Joachim (eds.) (2000), *Morphologie – Morphology: Ein Internationales Handbuch zur Flexion und Wortbildung – An International Handbook on Inflection and Word-Formation (Vol. 1)*, Berlin: Walter de Gruyter.

Bopp, Stephan (1993), *Computerimplementation der italienischen Flexions- und Wortbildungsmorphologie*, Hildesheim: Olms.

Carstairs-McCarthy, Andrew D. (1992), *Current Morphology,* London: Routledge.

Chandioux, John (1989), '10 Ans de METEO', in Abbou, André (ed.), *La Traduction Assistée par Ordinateur*, Paris: Daicadif, p. 169-175.

Cole, Ronald; Mariani, Joseph; Uszkoreit, Hans; Varile, Giovanni Battista; Zaenen, Annie; Zampolli, Antonio & Zue, Victor (eds.) (1997), *Survey of the State of the Art in Human Language Technology*, Cambridge: Cambridge University Press & Pisa: Giardini.

Corbin, Danielle (1987), *Morphologie dérivationelle et structuration du lexique*, Tübingen: Niemeyer (2 vol.).

Dale, Robert; Moisl, Hermann & Somers, Harold (eds.) (2000), *Handbook of Natural Language Processing*, New York: Dekker.

Deutsche Wortbildung 1: Kühnhold, Ingeburg & Wellmann, Hans (1973), *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache 1: Das Verb*, Düsseldorf: Schwann.

Deutsche Wortbildung 2: Wellmann, Hans (1975), *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache 2: Das Substantiv*, Düsseldorf: Schwann.

Deutsche Wortbildung 3: Kühnhold, Ingeburg; Putzer, Oskar & Wellmann, Hans (1978), *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache 3: Das Adjektiv*, Düsseldorf: Schwann.

Deutsche Wortbildung 4: Ortner, Lorelies; Bollhagen-Müller, Elgin; Ortner, Hanspeter; Wellmann, Hans; Pümpel-Mader, Maria & Gärtner, Hildegard (1991), *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache 4: Substantivkomposita*, Berlin: de Gruyter.

Deutsche Wortbildung 5: Pümpel-Mader, Maria; Gassner-Koch, Elsbeth & Wellmann, Hans (1992), *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache 5: Adjektivkomposita und Partizipialbildungen*, Berlin: de Gruyter.

Domenig, Marc (1989), *Word Manager: A system for the Specification, Use, and Maintenance of Morphological Knowledge*, Habilitationsschrift, Universität Zürich.

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

Eisenberg, Peter (1998), *Grundriss der deutschen Grammatik. Band 1: Das Wort*, Stuttgart: Metzler.

Endres-Niggemeyer, Brigitte (1998), *Summarizing Information*, Berlin: Springer.

Evans, Roger & Gazdar, Gerald (1996), 'DATR: A Language for Lexical Knowledge Representation', *Computational Linguistics* 22:167-216.

Evert, Stefan & Lüdeling, Anke (2001), 'Measuring morphological productivity: Is automatic preprocessing sufficient?', *Proceedings of Corpus Linguistics 2001*, Lancaster.

Fleischer, Wolfgang & Barz, Irmhild (1992), *Wortbildung der deutschen Gegenwartsprache*, Tübingen: Niemeyer.

Fuhrhop, Nanna (1998), *Grenzfälle morphologischer Einheiten*, Tübingen: Stauffenberg.

ten Hacken, Pius (1994), *Defining Morphology: A Principled Approach to Determining the Boundaries of Compounding, Derivation, and Inflection*, Hildesheim: Olms.

ten Hacken, Pius & Domenig, Marc (1996), 'Reusable Dictionaries for NLP: The Word Manager Approach', *Lexicology* 2:232-255.

ten Hacken, Pius (1998), 'Word Formation in Electronic Dictionaries', *Dictionaries* 19:158-187.

ten Hacken, Pius & Bopp, Stephan (1998), 'Separable Verbs in a Morphological Dictionary for German', in *Coling - ACL '98: Proceedings of the Conference*, Université de Montréal, p. 471-475.

ten Hacken, Pius (1999), 'Two Perspectives on the Reusability of Lexical Resources', *McGill Working Papers in Linguistics* 14:39-49.

ten Hacken, Pius & Tschichold, Cornelia (2001), 'Word Manager and CALL: Structured access to the lexicon as a tool for enriching learners' vocabulary', *ReCALL* 13: 121-131.

ten Hacken, Pius (2002), 'Word Formation and the Validation of Lexical Resources', to appear in the proceedings of LREC 2002 - Language Resources & Evaluation Conference, Las Palmas, 27 May - 2 June 2002.

ten Hacken, Pius & Smyk, Dorota (2002), 'Word Formation versus Etymology in Electronic Dictionaries', to appear in the proceedings of Euralex 2002, København, 13-17 August 2002.

Heid, Ulrich (2001), *DeKo: Derivations- und Kompositionsmorphologie, Zwischenbericht*, Technical report IMS, University of Stuttgart, available at http://www.ims.uni-stuttgart.de/projekte/DeKo/

Heid, Uli; Säuberlich, Bettina & Fitschen, Arne (2002), 'Using descriptive generalizations in the Acquisition of lexical data for a word formation analyzer', in *Proceedings of the Third International Conference on Language Resources and Engineering (LREC)*, Las Palmas, Gran Canaria.

Hockett, Charles F. (1954), 'Two Models of Grammatical Description', *Word* 10:210-231.

Holz, Dieter (1995), *Über das Entwerfen von Gebrauchssoftware: Lehren aus dem Entwurfsprozeß einer Arbeitsumgebung für einen Lexikographen*, unpublished Ph.D. Dissertation, Universität Basel.

Hsiung, Alain (1995), *Lexicon Acquisition through High-Level Rule Compilation*, Hildesheim: Olms.

Karlsson, Fred & Karttunen, Lauri (1997), 'Sub-Sentential Processing', in Cole et al. (eds.), p. 96-100.

Kiraz, George A. (2001), *Computational Nonlinear Morphology With Emphasis on Semitic Languages*, Cambridge: Cambridge University Press.

Koskenniemi, Kimmo (1983), *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*, University of Helsinki, Department of General Linguistics Publications No. 11.

Lieber, Rochelle (1992), *Deconstructing Morphology: Word Formation in Syntactic Theory*, Chicago: University of Chicago Press.

Lüdeling, Anke & Fitschen, Arne (2002), 'An integrated lexicon for the analysis of complex words' to appear in *Proceedings of EURALEX 2002*, Copenhagen.

Lüdeling, Anke; Schmid, Tanja & Kiokpasoglou, Sawwas (2002), 'On neoclassical word formation in German', to appear in *Yearbook of Morphology 2001*.

Manning, Christopher & Schütze, Hinrich (1999), *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press.

Matthews, Peter H. (1974), *Morphology: An Introduction to the Theory of Word Structure*, Cambridge: Cambridge University Press.

Pedrazzini, Sandro (1994), *Phrase Manager: A system for Phrasal and Idiomatic Dictionaries*, Hildesheim: Olms.

Pedrazzini, Sandro & ten Hacken, Pius (1998), 'Centralized Lexeme Management and Distributed Dictionary Use in Word Manager™', in Schröder, Bernhard; Lenders, Winfried; Hess, Wolfgang & Portele, Thomas (eds.), *Computers, Linguistics and Phonetics between Language and Speech, Proceedings of the 4th Conference on NLP, Konvens'98, Bonn, Germany*, Frankfurt am Main: Lang, p. 365-370.

Pedrazzini, Sandro (1999), 'The Finite State Automata's Design Patterns', in Champarnaud, Jean-Marc; Maurel, Denis & Ziadi, Djelloud (eds.), *Automata Implementation, Third International Workshop on Implementing Automata, WIA'98, Rouen, France*, Berlin: Springer, p. 213-219.

Plag, Ingo (1999), *Morphological Productivity: Structural Constraints in English Derivation*, Berlin: Mouton de Gruyter.

Ritchie, Graeme (1987), 'The Lexicon', in Whitelock et al. (eds.), p. 225-256.

Säuberlich, Bettina (2001), *Aufbau und Regelformat von DeKo*, Technical report IMS, available at http://www.ims.uni-stuttgart.de/projekte/DeKo/

Schiller, Anne (1996), 'Deutsche Flexions und Kompositionsmorphologie mit PC-KIMMO', in Hausser, Roland (ed), *Linguistische Verifikation. Dokumentation zur ersten Morpholympics 1994*, Tübingen: Niemeyer.

Schmid, Helmut (1994), 'Probabilistic part-of-speech tagging using decision trees', in *International Conference on New Methods in Language Processing*, Manchester, p. 44-49.

Schmid, Tanja; Lüdeling, Anke; Säuberlich, Bettina; Heid, Ulrich and Möbius, Bernd (2001), 'DeKo: Ein System zur Analyse komplexer Wörter', in *GLDV - Jahrestagung 2001*, p. 49-57.

Spencer, Andrew (1991), *Morphological Theory; An Introduction to Word Structure in Generative Grammar*, Oxford: Blackwell.

Spencer, Andrew & Zwicky, Arnold M. (eds.) (1998), *The Handbook of Morphology*, Oxford: Blackwell.

Sproat, Richard (1992), *Morphology and Computation*, Cambridge (Mass.): MIT Press.

Sproat, Richard (2000a), 'Lexical Analysis', in Dale et al. (eds.), p. 37-57.

Sproat, Richard (2000b), *Lextools. A toolkit for finite-state linguistic analysis,* Technical report, available at http://www.research.att.com/sw/tools/lextools

Trost, Harald (1993), 'Coping With Derivation in a Morphological Component', in *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, p. 368-376.

Tschichold, Cornelia (2000), *Multi-Word Units in Natural Language Processing*, Hildesheim: Olms.

Volk, Martin & Clematide, Simon (2001), 'Learn-Filter-Apply-Forget. Mixed Approaches to Named Entity Recognition', in *Proceedings of the 6th International Workshop on Applications of Natural Language for Information Systems*, Madrid.

Whitelock, Pete; McGee Wood, Mary; Somers, Harold L.; Johnson, Rod & Bennett, Paul (eds.) (1987), *Linguistic Theory and Computer Applications*, London: Academic Press.

Zappatore, Daniela & ten Hacken, Pius (2000), 'Word Manager and Banking Terminology: Industrial Application of a General System', in Heid, Ulrich; Evert, Stefan; Lehmann, Egbert & Rohrer, Christian (eds.), *Proceedings of the Ninth Euralex International Congress, Euralex 2000*, Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, p. 325-335.