Sofia University

Faculty of Mathematics and Informatics

Department of Information Technologies

Recognition and Morphological Classification of Unknown Words for German

Preslav Nakov, fn42408, Informatics

Advisor: Galia Angelova, Ph.D., *Associate Professor,*
Linguistic Modelling Department,
Central Laboratory for Parallel Processing,
Bulgarian Academy of Sciences

Sofia

July 2001

Abstract

A system for recognition and morphological classification of unknown words for German is described. The System takes raw text as input and outputs list of the unknown nouns together with hypothesis about their possible morphological class and stem. The morphological classes used uniquely identify the word gender and the inflection endings it takes when changes by case and number. The System exploits both global (ending guessing rules, maximum likelihood estimations, word frequency statistics) and local information (surrounding context) as well as morphological properties (compounding, inflection, affixes) and external knowledge (specially designed lexicons, German grammar information etc.). The problem is solved as a sequence of subtasks including: unknown words identification, noun identification, inflected forms of the same word recognition and grouping (they must share the same stem), compounds splitting, morphological stem analysis, stem hypothesis for each group of inflected forms, and finally — production of ranked list of hypotheses about the possible morphological class for each group of words. The System is a kind of tool for lexical acquisition: it identifies, derives some properties and classifies unknown words from a raw text. Only nouns are currently considered but the approach can be successfully applied to other parts-of-speech as well as to other inflexional languages.

1 Introduction

1.1 The problem

The problem of the unknown words is a general problem for every natural language processing (NLP) system. No matter how big *lexicon* (machine-readable dictionary) it has, there will be always unknown words present. New words are constantly added to the language while others are no longer used. The natural language is dynamic in its nature and it is impossible to design so huge dictionary that will contain all the words that could appear in a real-life text: new words are constantly added to the language, other words get less frequent and are dropped out, while some of the existing ones lose, change or obtain new meaning. Even if one manages to build a complete dictionary it will be no longer valid in only few days since new words will inevitably appear.

The two major sources of new words for every natural language are the *proper nouns* and the *foreign words*. They cause a big problem to the NLP applications because they are uncontrollable and theoretically unlimited. Nobody can predict all the foreign words that could enter the language. Anyway, while one could hope that one day every language would have an extensive dictionary, this is much more unlikely in what about the proper nouns. It is impossible to know all the names of places, persons or companies all over the world.

Another important source of new words is the word form generation process, which directly influences the lexical richness of the language. There are three major linguistic phenomena in this respect: *inflexion*, *derivation* and *compounding*.

The inflexion is very unlikely to produce an unknown word form unless the base form is unknown as well. A known word would hardly produce a new unknown word through inflexion. The inflexion process is more or less standardised for each language and the inflected forms for each known word are usually known as well. The inflection rules can differ for the different words according to their gender and/or ending etc. Anyway, they are quite stable and the language tends to have a limited set of morphological classes that cover all the words with possibly only very few exceptions. The new words that enter the language for most of the cases follow these general rules.

The derivation process is more powerful. Unlike the inflexion the derivation produces words that have possibly different part-of-speech (POS). A word obtained through derivation is a new word and not just a form of the base. The words obtained through derivation would be listed in a general-purpose human-readable dictionary as separate entries while the inflected forms are not present there. The derivation process is more powerful and more likely to produce new words. Anyway, the production of new words is not very likely unless the base form is a new word.

Both inflexion and derivation are standard processes for all European languages and generate large amount of word forms. The power of these processes differs in the different languages. The Slavonic and Roman languages, for example, are highly inflexional while English is poor in inflexions. Anyway, even in English a considerable amount of words are due to inflexions, while this is not the case with the compounding. The compounding is the process of concatenation of two or more words to form a new one with possibly new meaning. Almost all the European languages produce only a very limited amount of compounds, but this is not valid for German. The compounding process is very powerful in German since it is derivative. The word forms a base form can produce through both inflexion and derivation are limited and can be predicted in advance: all the rules are standard. At the same time a German word can enter in virtually unlimited amount of different compounds with other words. The process is very powerful not only theoretically, but also in practice: a large part of the unknown words in German are due to the compounding process.

Remark

There is another important source of unknown words in the real texts due to incorrectly written words. And especially for German there is another recent source of new word forms: the orthographic reform, which is not widely accepted. Since some parts of the population keep using the old orthography and other — the new one, this resulted in variety of new word forms. ?

1.2 The system

Our goal is the design and implementation of a system for identification and morphological classification of unknown words for German. The present system is limited to nouns only but the same approach would work for the other open POS: verbs, adjectives and adverbs.

The System accepts raw text as input and produces a list of unknown words together with hypotheses for their *stem* and *morphological class*. The stem is the common part shared by all inflected forms of the base while the morphological class describes both the word gender and the inflexion rules the word follows when changes by case and number. The notions of stem and morphological class will be explained in more details below. The stem and the morphological class together determine in an unambiguous way all the word forms that could be obtained through inflexion.

The System solves the problem as a sequence of subtasks including: unknown words identification, noun identification, inflected forms of the same word recognition and grouping (they must share the same stem), compounds splitting, morphological stem analysis, stem hypothesis for each group of inflected forms, and finally — production of ranked list of hypotheses about the possible morphological class for each group of words. This is a complex several-stage process, which exploits:

- **local context** (surrounding context: articles, prepositions, pronouns)
- **global context** (ending guessing rules, maximum likelihood estimations, word frequency statistics)
- **morphology** (compounding, inflection, affixes)
- **external sources** (specially designed lexicons, German grammar information etc.)

1.3 Areas of application

What the System is and what is not. The System is a kind of tool for lexical acquisition: it identifies, derives some properties and classifies unknown words from a raw text. It could be used as a tool for automatic dictionary extension with new words.

1.3.1 POS guesser

The System is not a POS guesser in its traditional meaning. The purpose of the POS guesser is to make a hypothesis about the possible POS for an unknown word looking at its graphemic form and possibly in a lexicon. Our System is not restricted to the local context and considers all the word occurrences. We are not interested in the exact POS of a word but just in whether it is a noun. And once we know it is a noun we do not stop there but we continue the work trying to identify other inflectional forms of the same word and derive a hypothesis for its morphological class (this includes the gender identification). Anyway, the System could be seen as kind of morphological class guesser.

1.3.2 Morphological analyser

The System is not a pure morphological analyser although it can be used as such, since it outputs the morphological information available for the known words just like a morphological analyser does. Anyway, it works at global level, which means it does not try to disambiguate between the possible lexical forms of a specific word token. We are not interested in a particular word token in a context but in the word type the word token is instance of. The morphological analysers usually output all the possible morphological information. But sometimes try to disambiguate between the possible morphological forms the observed graphemic form is an instance of. In the later case they act in combination with a POS tagger and the morphological analyser works as an extended POS tagger, which adds morphological information (gender, case and number) to the POS tags. The morphological analysers usually have some local strategies to deal with unknown words but this is not a central task for them and they often use only simple heuristics.

1.3.3 Stemmer

The System is not a stemmer in the classic meaning of that word, although it outputs the stems for the known nouns and makes hypothesis for the possible stems of the unknown nouns. What is important here is that the stem we produce groups together the inflected word forms only. But the classic notion of stemming as used in information retrieval conflates both inflectional and derivational forms. Thus, *generate*, *generator* are grouped together with a classic stemmer but not with our System.

1.3.4 Lemmatiser

The System is not a lemmatiser but could be used as such since it outputs both the stem and the morphological class for each word. Usually the stem and the lemma are the same but there are some exceptions. Anyway, given the morphological class and stem the lemma identification is straightforward.

1.3.5 Compound analyser

The compound analysis is a substantial part of the System although this is not a central task. Anyway, every unknown word is analysed as a potential compound. In case there is at least one legal way to split it, we recognise it as a compound. But we are not interested in the actual compound splitting and we output only the last part of the splitting. In case there is more than one possibility for the last part we output all the possibilities. But we never output the splitting of the first part, although we always obtain it internally.

2 Terminology used

Notion	Meaning
POS	Part-Of-Speech
Word	In the remaining text we will try use *word* in its most general sense.
Noun	For most of the cases *noun* will be used in the sense of a specific part of speech. Unless the context does not permit this, the word *noun* will be always supposed to include the meanings of both *common* and *proper* nouns. Otherwise it will mean just common noun.
Word type **Ending**	Group of tokens with exactly the same graphemic form. In general the last few letters a word type ends with. Anyway, for German his notion is extended to account for the umlauts and *ß* alternations.
Base form	This is the singular nominative form of the German noun.
Stem	The stem is the string shared by all inflected forms of a noun when it changes by case and number. The changes caused by umlauts and words ending by “ß” are not considered to change the stem while in fact they do so.

Table 1. Terminology used.

Remarks:

1. In fact the base form and stem differ for only few cases. This happens for words from four of the classes: *m11* (e.g. *der Organismus*, stem *Organism*), *f15a* (e.g. *die Firma,* stem *Firm*), *n28* (e.g. *das Datum*, stem *Dat*), *n28a* (e.g. *das Drama,* stem *Dram*).
2. For others there are two possibilities for the base form and since only one of them is the stem, the other one leads to difference: *m7* (e.g. *der Bekannte*/*Bekannter*, stem *Bekannte*), *n26* (e.g. *das Junge/Junges*, stem *Junge*).
3. Some of the morphological classes change a short stem vowel to umlaut when forming the plural inflected forms: *m2*, *m3*, *m5*, *f14*, *f14a*, *n20a*, *n22* and *n23a*.
4. Each word ending by *ß* changes to *ss* in some of the inflected forms and this happens for all morphological classes (e.g. *der Fuß* → *Fusse*). ?

3 Related work

As we saw above the System's task is more or less related to several classic NLP tasks. Anyway, it is obvious that the nearest task is the one of morphological analysis, while other tasks like stemming are much more dissimilar. Below we consider briefly the related work and then several important systems for morphological analysis are described in more details.

(Deshler, Ellis & Lenz, 1996) advice useful strategies and methods for adolescents with learning disabilities for coping with unknown words. These techniques are particularly useful for NLP: An unknown word could be recognised through: a) context analysis, b) semantic analysis, c) structural analysis, d) morphological analysis and e) external sources (e.g. dictionary).

Koskenniemi proposes a language independent model for both morphological analysis and generation called *two-level morphology* and based on finite-state automata. It lies behind several systems including *KIMMO* (Koskenniemi, 1983a, 1983b) and *GERTWOL* (Haapalainen and Majorin, 1994). A similar approach based on augmented two-level morphology is described by (Trost, 1991, 1985). Useful sets of finite state utilities are implemented by (Daciuk, 1997). Finkler and Neumann follow a different approach using *n*-ary tries in their system *MORPHIX* (see Finkler and Neumann, 1988; Finkler and Lutzky, 1996). (Lorenz, 1996) developped *Deutsche Malaga-Morphologie* as a system for the automatic word form recognition for German based on *Left-Associative Grammar* using the *Malaga* system. (Karp et al., 1992) present a freely available morphological analyser for English with an extensive lexicon. Under the *MULTEXT* project (Armstrong et al., 1995; Petitpierre and Russell, 1995) provided morphological analysers and other linguistic tools for six different European languages.

(Neumann and Mazzini, 1999; Neumann et al., 1997) consider the problem of compound analysis by means of longest matching substrings found in the lexicon. (Adda-Decker & Adda, 2000) propose general rules for morpheme boundary identification. These are hypothesised after the occurrence of sequences such as: *-ungs*, *-hafts*, *-lings*, *-tions*, *-heits*. The problem of German compounds is considered in depth by (Goldsmith and Reutter, 1998; Lezius, 2000; Ulmann, 1995). (Hietsch, 1984) concentrates on the function of the second part of a German compound.

(Kupiec, 1992) uses pre-specified suffixes and then learns statistically the POS predictions for unknown word guessing. The XEROX tagger comes with a list of built-in ending guessing rules (Cutting et al., 1992). In addition to the ending (Weischedel et al., 1993) considers the capitalisation feature in order to guess the POS. (Thede & Harper, 1997) and (Thede, 1997) consider the statistical methods for unknown words tagging using contextual information, word endings, entropy and open-class smoothing. Similar approach is presented in (Schmid, 1995). (Rapp, 1996) derives useful German suffix frequencies. A revolutionary approach has been proposed by Brill (Brill 1995, 1999). He builds more linguistically motivated rules by means of tagged corpus and a lexicon. He does not look at the affixes only but optionally check their POS class in a lexicon. The prediction is trained from a tagged corpus. Mikheev proposes a similar approach that estimates the rule predictions from a raw text (Mikheev 1997, 1996a, 1996b, 1996c). Daciuk observes that the rules thus created could be implemented as finite state transducers in order to speed up the process (Daciuk, 1997).

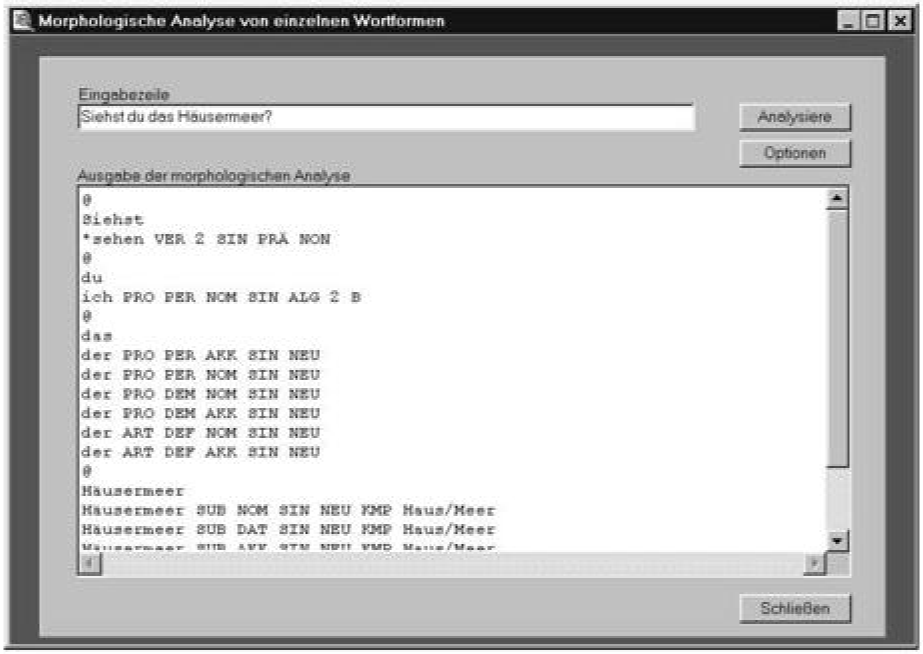
Schone and Jurafsky propose the usage of *Latent Semantic Analysis* for a knowledge-free morphology induction (Schone and Jurafsky, 2000). Goldsmith proposes a *Minimum Description Length analysis* to model unsupervised learning of the morphology of European languages, using corpora ranging in sizes from 5,000 word to 500,000 words. (Goldsmith, 2000). Kazakov uses *genetic algorithms* (Kazakov, 1997). (Goldsmith, 2000) cuts the words in exactly one place and hypothesises the stem and suffix. (DeJean, 1998) cuts the word if the number of distinct letters following a pre-specified letter sequence surpasses a threshold using an approach similar to the one proposed by (Hafer & Weiss, 1974). (Gaussier, 1999) tries to find derivational morphology in a lexicon by a *p*-similarity based splitting. (Jacquemin, 1997) focuses on learning morphological processes. (Van den Bosch & Daelemans, 1999) propose a memory-based apporach mapping directly from letters in context to rich categories that encode morphological boundaries, syntactic class labels, and spelling changes. (Viegas et al., 1996) use derivational lexical rules to extend a Spanish lexicon. (Yarowsky & Wicentowski, 2000) present a corpus based approach for morphological analysis of both regular and irregular forms based on 4 original models including: relative corpus frequency, context similarity, weighted string similarity and incremental retraining of inflectional transduction probabilities. Another approach exploiting capitalisation, as well as both fixed and variable suffix is proposed in (Cucerzan & Yarowsky, 2000).

(Lovins, 1968) and (Porter, 1980) devised the classic manually build stemming algorithms. (Hull, 1996), (Harman, 1991), (Kraaij, 1996) and (Krovetz, 1993) discuss the impact of the stemming algorithms. (Popovic & Willett, 1992) consider the application of stemming to Slovene. (Xu & Croft, 1998) propose a corpus based stemming algorithm.

(Krovetz, 1993) has shown that the correct recognition of the morphological variants is of particular importance for information retrieval (IR). (Hoch, 1994) demonstrates the usage of the morphological system MORPHIX (see Finkler and Neumann, 1988; Finkler and Lutzky, 1996) for IR terms analysis in his INFOCLAS system for statistical information retrieval. Adda-Decker and Adda consider the morphological analysis application to automatic speech recognition for German (Adda-Decker & Adda, 2000). (Weischedel et al., 1993) study the impact of the unknown words on the effectiveness of the application of probabilistic methods for POS tagging and conclude that the morphological information could improve the results by a factor of 5.

3.1 Morphy

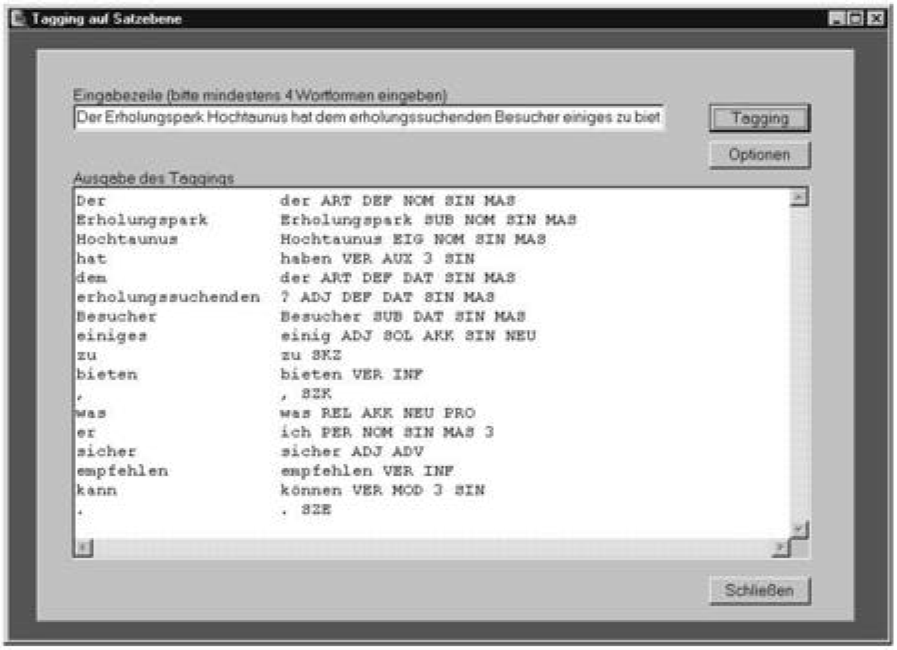
The morphological system Morphy is developed by Wolfgang Lezius as an integrated tool for German morphology, part-of-speech tagging and context-sensitive lemmatisation. The output of the morphological analysis is usually highly ambiguous (see Figure 1). Syntactic ambiguities can be resolved with a standard statistical part-of-speech tagger. By using the output of the tagger, the lemmatiser can determine the correct root even for ambiguous word forms. (see Figure 2) The package is developed in Delphi, runs under Windows and can be downloaded from the World Wide Web at http://www-psycho.uni-paderborn.de/lezius/. Morphy can generate output in variety of formats including HTML, SGML, XML, plain text etc. The annotated output of Morphy can be imported directly into the *Tatoe* corpus query tool (Rostek & Alexa, 1998), which could be freely downloaded at http://www.darmstadt.gmd.de/~rostek/tatoe.htm.


Morphologische Analyse von einzelnen Wortformen
Eingabezeile
Analysiere
Optionen
Ausgabe der morphologischen Analyse
Schließen

Siehst du das Häusermeer?

```
@
Siehst
*sehen VER 2 SIN PRÄ NON
@
du
ich PRO PER NOM SIN ALG 2 B
@
das
der PRO PER AKK SIN NEU
der PRO PER NOM SIN NEU
der PRO DEM NOM SIN NEU
der PRO DEM AKK SIN NEU
der ART DEF NOM SIN NEU
der ART DEF AKK SIN NEU
@
Häusermeer
Häusermeer SUB NOM SIN NEU KMP Haus/Meer
Häusermeer SUB DAT SIN NEU KMP Haus/Meer
Häusermeer SUB AKK SIN NEU KMP Haus/Meer
```

Figure 1. ***Morphy:* Morphological analysis of an example sentence.**



Der Erholungspark Hochtaunus hat dem erholungssuchenden Besucher einiges zu biet

```
Der                  der ART DEF NOM SIN MAS
Erholungspark        Erholungspark SUB NOM SIN MAS
Hochtaunus           Hochtaunus EIG NOM SIN MAS
hat                  haben VER AUX 3 SIN
dem                  der ART DEF DAT SIN MAS
erholungssuchenden   ? ADJ DEF DAT SIN MAS
Besucher             Besucher SUB DAT SIN MAS
einiges              einig ADJ SOL AKK SIN NEU
zu                   zu SKZ
bieten               bieten VER INF
,                    , SZK
was                  was REL AKK NEU PRO
er                   ich PER NOM SIN MAS 3
sicher               sicher ADJ ADV
empfehlen            empfehlen VER INF
kann                 können VER MOD 3 SIN
.                    . SZE
```

Figure 2. *Morphy:* Tagging and lemmatisation of an example sentence.

A basic resource for most of the NLP applications is the lexicon. Unfortunately, these resources are not widely available and their manual construction is very hard and time consuming. In our present work we use the Morphy lexicon as a base for our own lexicons construction since it offers a free lexicon of 50,500 stems and 324,000 different word forms. The package is able to export parts or the whole lexicon using two tag sets — a small (51 tags) and a large one (about 1000 tags).

3.2 PC-KIMMO

PC-KIMMO is a PC version of the programme KIMMO, originally created by Prof. Kimmo Koskenniemi in 1983. The programme is based on two-level morphology and its purpose is the generation and/or recognition of words. The two-level model of word structure is a model in which a word is represented as a correspondence between its lexical level form and its surface level form. (see Koskenniemi, 1993, 1984, 1983a, 1983b; Antworth, 1990; Karttunen, 1983; Sproat, 1991)

PC-KIMMO is language independent and expects that the user provides it a description of a language, which consists of two files:

1. a *rules file***,** which specifies the alphabet and the phonological (or spelling) rules, and
2. a *lexicon file***,** which lists lexical items (words and morphemes) and their glosses, and encodes morphotactic constraints.

The theoretical model of phonology embodied in PC-KIMMO is called two-level phonology. In the two-level approach, phonological alternations are treated as direct correspondences between the underlying (or lexical) representation of words and their realisation on the surface level. One character of the lexical level corresponds to one character (possibly a null character) of the surface level. This makes both analysis and generation of word forms possible with the same morphological description. The two-level model has been used for describing approximately 30 languages.

For example, to account for the rules of English spelling, the surface form *spies* must be related to its lexical form *`spy+s* as follows (where ` indicates stress, + indicates a morpheme boundary, and 0 indicates a null element):

```
Lexical Representation:   ` s p y + 0 s
Surface Representation:   0 s p i 0 e s
```

Rules must be written to account for the special correspondences `:0, y:i, +:0, and 0:e. For example, the two-level rule for the y:i correspondence looks like this (somewhat simplified):

```
y:i => @:C___+:0
```

Notice that the environment of the rule is also specified as a string of two-level correspondences. Because two-level rules have access to both underlying and surface environments, interactions among rules can be handled without using sequential rule ordering. All of the rules in a two-level description are applied simultaneously, thus avoiding the creation of intermediate levels of derivation (an artefact of sequentially applied rules).

```
        +-----------+        +-----------+
        |  RULES    |        |  LEXICON  |
        +----+------+        +------+----+
             |-------+      +-------|
                     |      |
                     v      v
Surface Form:  +------------------+      Lexical Form:
 spies ------->|    Recognizer    |----> `spy+s
               +----+-------------+      [N(spy)+PLURAL]
                    |
                    v
               +------------------+
 spies <-------|    Generator     |<----- `spy+s
               +------------------+
```

Figure 3. PC -KIMMO: main components**.**

The two functional components of PC-KIMMO are the generator and the recogniser. The generator accepts as input a lexical form, applies the phonological rules, and returns the corresponding surface form. It does not use the lexicon. The recogniser accepts as input a surface form, applies the phonological rules, consults the lexicon, and returns the corresponding lexical form with its gloss (see Figure 3). The rules and the lexicon are implemented computationally using finite state machines. The PC-KIMMO system can be run in both interactive and batch mode and provides a useful set of debugging facilities including automatic comparison of the results to the correct ones previously supplied by the user.

Because the PC-KIMMO user shell is intended to facilitate development of a description, its data-processing capabilities are limited. The primitive PC-KIMMO functions (including load rules, load lexicon, generate, recognise) are available as a source code library that can be included in another program. This means that the users can develop and debug a two-level description using the PC-KIMMO shell and then link PC-KIMMO's functions into their own programs. The programme is available at http://www.sil.org/pckimmo/about_pc -kimmo.html

3.3 GERTWOL

GERTWOL (Haapalainen and Majorin, 1994) is a commercial language-independent system based on the two-level model and on the ideas of Prof. Kimmo Koskenniemi used in the KIMMO system (Koskenniemi, 1983a, 1983b). The system concentrates on the automatic recognition and morphological analysis of German word forms. GERTWOL has been tested on various text corpora: newspaper articles, legal documents, weather forecasts, literary texts and business reports and is reported to achieve 99% coverage for correctly spelled texts and more than 98% for unrestricted texts (September 1994).

The basic lexicon of GERTWOL is the complete material of the Collins German Dictionary supplemented with more than 6,300 common nouns and 11,000 proper nouns. The approximate numbers of lexemes for different parts of speech are as follows: 11,000 adjectives, 2,000 adverbs, 400 interjections, 50,000 common nouns, 6,500 verbs, 12,000 proper nouns and 1,700 abbreviations. Adding conjunctions, pronouns, articles and prepositions results in a lexicon of approximately 85,000 words. This number is considerably increased by an extensive derivational morphology and a complete mechanism for compounding. More information about the GERTWOL lexicon can be found at http://www.uni-koblenz.de/~gtu/GERTWOLLex.html. Figure 4 shows the results of the analysis of the German sentence "*GERTWOL ist ein System zur automatischen Wortformerkennung deutscher Wörter*".

Sample Analysis

GERTWOL ist ein System zur automatischen Wortformerkennung deutscher Wörter.

```
GERTWOL
    "*g*e*r*t*w*o*l"  ABK S EIGEN

ist
    "sein"  V IND PRÄS SG3

ein
    "ein"  PRÄF
    "ein"  ADV
    "ein"  NUM KARD
    "einen"  V IMP PRÄS GESPROCHEN SG2
    "ein"  ART INDEF SG AKK NEUTR
    "ein"  ART INDEF SG NOM NEUTR
    "ein"  ART INDEF SG NOM MASK

System
     "*system"  S NEUTR SG DAT
     "*system"  S NEUTR SG AKK
     "*system"  S NEUTR SG NOM

zur
    "zu-die"  PRÄP ART DEF SG DAT FEM
```

```
automatischen
    "automat~isch"  A POS PL GEN
    "automat~isch"  A POS PL AKK
    "automat~isch"  A POS PL NOM
    "automat~isch"  A POS SG GEN FEM
    "automat~isch"  A POS SG DAT FEM
    "automat~isch"  A POS SG DAT NEUTR
    "automat~isch"  A POS SG DAT MASK
    "automat~isch"  A POS PL DAT
    "automat~isch"  A POS SG GEN NEUTR
    "automat~isch"  A POS SG GEN MASK
    "automat~isch"  A POS SG AKK MASK

Wortformerkennung
    "*wort#form~er#kenn~un g"  S FEM SG GEN
    "*wort#form~er#kenn~ung"  S FEM SG DAT
    "*wort#form~er#kenn~ung"  S FEM SG AKK
    "*wort#form~er#kenn~ung"  S FEM SG NOM
    "*wort#form#er|kenn~ung"  S FEM SG GEN
    "*wort#form#er|kenn~ung"  S FEM SG DAT
    "*wort#form#er|kenn~ung"  S FEM SG AKK
    "*wort#form#er|kenn~ung"  S FEM SG NOM

deutscher
    "deutsch"  A KOMP
    "deutsch"  A POS PL GEN
    "deutsch"  A POS SG GEN FEM
    "deutsch"  A POS SG DAT FEM
    "deutsch"  A POS SG NOM MASK

Wörter
    "*wort"  S NEUTR PL GEN
    "*wort"  S NEUTR PL AKK
    "*wort"  S NEUTR PL NOM

--punkt
    ""  PUNKT
```

Figure 4. GERTWOL: Analysis of a sample sentence.

3.4 QuickTag

Quicktag is a COM component for Win32 that can efficiently tag (identify the possible grammatical categories of words), lemmatise (identify the root form of words), disambiguate (indicate the actual grammatical category of words) and extract noun phrases from English text. The programme is written by Michael Decary from Cogilex R&D Inc. and runs under both Windows and Linux. Figure 5 shows the system at work.

ORIGINAL SENTENCE:

```
   Functional changes are early indicators of growth in clonal development of the
hematopoietic system but they equally indicate signalling for specific actions of
differentiated cells
```

QUICKTAG ANALYSIS:

```
Number of Words: 25
   Word              Lemma             P.O.S.
   Functional        functional        Adj
   changes           change            N(Plural)
   are               be                V(Pres)
   early             early             Adv
   indicators        indicator         N(Plural)
   of                of                Prep(VComp)
   growth            growth            N
   in                in                Prep(VComp)
```

```
clonal            clonal            Adj
development       development       N
of                of                Prep(VComp)
the               the               Det(DefiniteArticle)
hematopoietic     hematopoietic     Adj
system            system            N
but               but               Conj
they              they              Pro(definite)
equally           equally           Adv
indicate          indicate          V
signalling        signal            Ing
for               for               Prep(VComp)
specific          specific          Adj
actions           action            N(Plural)
of                of                Prep(VComp)
differentiated    differentiated    Adj
cells             cell              N(Plural)
```

Figure 5. QuickTag system work demonstration.

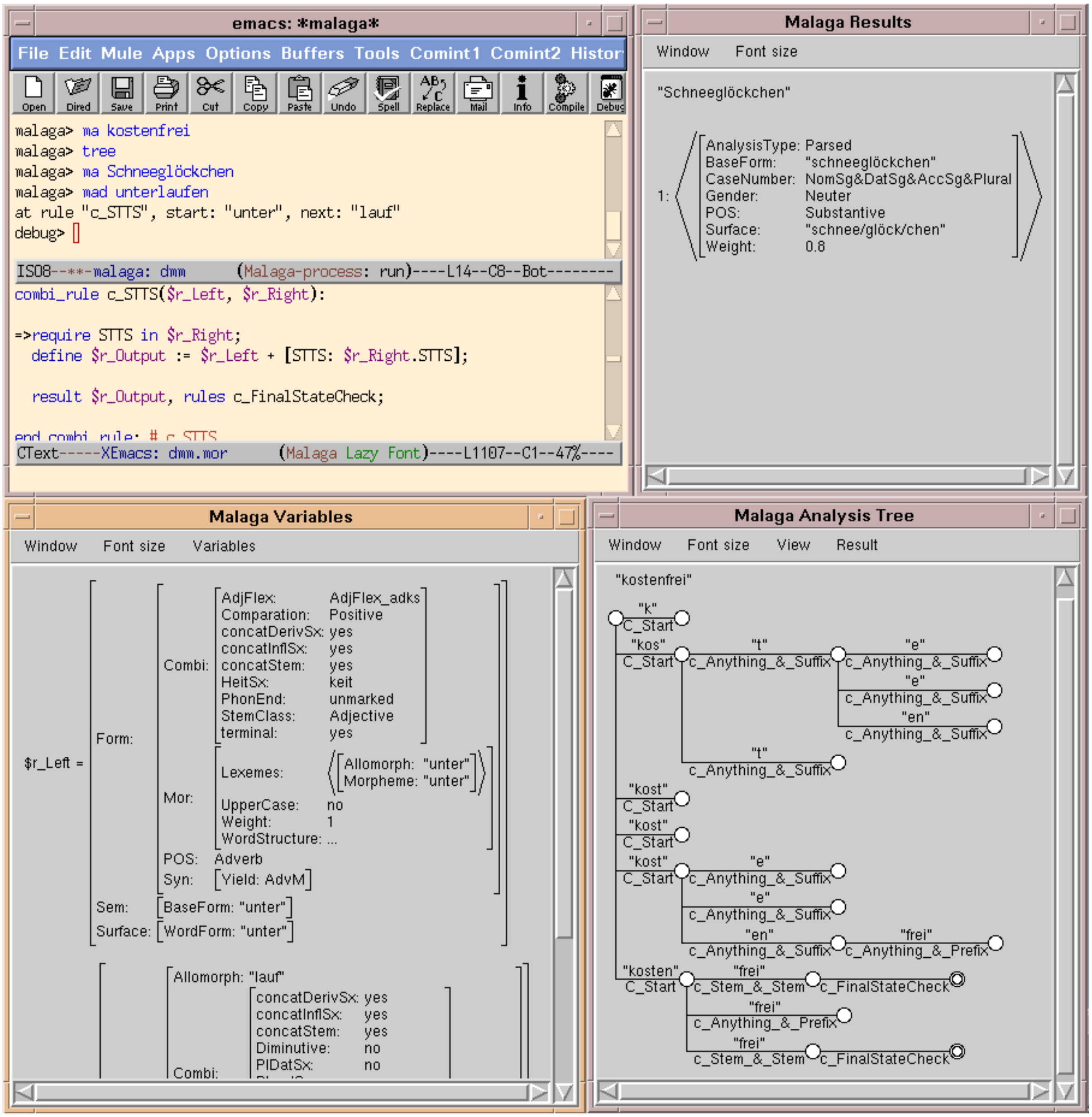
The system is patented by Cogilex. In addition to QuickTag the parsing component QuickParse is provided. Both QuickTag and QuickParse are implemented as generic libraries that can be adapted to user specific needs or could be used as part of a complete NLP solution. The system can be purchased at http://www.cogilex.com/products.htm. It is possible to try it on-line as well at: http://www.cogilex.com/online.asp.

3.5 Deutsche Malaga-Morphologie

Malaga has been developed by Bjoern Beutel, at the Computational Linguistics Department of the *University Erlangen-Nurnberg* (CLUE), as a software package for linguistic applications within the framework of *Left-Associative Grammar* (LAG). In contrast to Phrase Structure Grammars, which are based on the principle of possible substitutions, LAGs are based on the principle of *possible continuations*. The input is analysed *left-associatively* (left to right in the case of Western scripts, more generally: in writing direction). Analysis is *time-linear* and *surface-compositional*, which means that the input segments are concatenated in order of their occurrence (left to right) and each rule application is necessarily linked with reading exactly one input segment.

Malaga contains a programming language for the modelling of morphology and syntax grammars. Rule and lexicon compilers, which translate grammar components written by developers into a binary format as well as a run time component that can analyse word forms or whole texts are available. Figure 6 shows a sample screen from a Malaga session under UNIX. The package contains some example grammars for formal languages, and a German toy syntax grammar. Full-grown morphology components for the German, Italian, Korean and English language have been developed.

Deutsche Malaga-Morphologie (DMM) is developed by Lorenz as a system for the automatic word form recognition of German. It is based on Malaga and concentrates on *categorisation* (assigning grammatical categories like part of speech, case, gender, number, person, tense etc., to a word form), *lemmatisation* (assigning a base form to a wordform) and *segmentation* (identification of the morphemes a word form is composed of). DMM works with a base form lexicon with about 50,000 entries, consisting of: 20,400 nouns, 11,200 adjectives, 10,900 proper nouns, 6,200 verbs. The rest are function words (determiners, prepositions, etc.), inflectional endings, prefixes, linking morphemes, etc. By means of special rules 67,000 allomorphs are generated from these 50,000 entries. The allomorphs are then concatenated to word forms by the run-time component. (see Lorenz, 1996).


emacs: *malaga*
File Edit Mule Apps Options Buffers Tools Comint1 Comint2 Histor
Open Dired Save Print Cut Copy Paste Undo Spell Replace Mail Info Compile Debug
Malaga Results
Window Font size
Malaga Variables
Window Font size Variables
Malaga Analysis Tree
Window Font size View Result

```
malaga> ma kostenfrei
malaga> tree
malaga> ma Schneeglöckchen
malaga> mad unterlaufen
at rule "c_STTS", start: "unter", next: "lauf"
debug>
ISO8--**-malaga: dmm        (Malaga-process: run)----L14--C8--Bot--------
combi_rule c_STTS($r_Left, $r_Right):

=>require STTS in $r_Right;
  define $r_Output := $r_Left + [STTS: $r_Right.STTS];

  result $r_Output, rules c_FinalStateCheck;

end combi_rule; # c_STTS
CText-----XEmacs: dmm.mor       (Malaga Lazy Font)----L1107--C1--47%----
```

```
"Schneeglöckchen"

    /[AnalysisType: Parsed                          ]\
    |[BaseForm:     "schneeglöckchen"               ]|
    |[CaseNumber:   NomSg&DatSg&AccSg&Plural        ]|
1: <[Gender:       Neuter                          ]>
    |[POS:          Substantive                     ]|
    |[Surface:      "schnee/glöck/chen"             ]|
    \[Weight:       0.8                             ]/
```

```
           [      [       [AdjFlex:         AdjFlex_adks]               ]]
           [      [       [Comparation:     Positive    ]               ]]
           [      [       [concatDerivSx:   yes         ]               ]]
           [      [       [concatInflSx:    yes         ]               ]]
           [      [Combi: [concatStem:      yes         ]               ]]
           [      [       [HeitSx:          keit        ]               ]]
           [      [       [PhonEnd:         unmarked    ]               ]]
           [      [       [StemClass:       Adjective   ]               ]]
           [      [       [terminal:        yes         ]               ]]
           [Form: [                                                     ]]
$r_Left =  [      [       [Lexemes:  <[Allomorph: "unter"]>]            ]]
           [      [       [          <[Morpheme:  "unter"]>]            ]]
           [      [Mor:   [UpperCase:     no              ]            ]]
           [      [       [Weight:        1               ]            ]]
           [      [       [WordStructure: ...             ]            ]]
           [      [POS:   Adverb                                        ]]
           [      [Syn:   [Yield: AdvM]                                 ]]
           [Sem:     [BaseForm: "unter"]                                 ]
           [Surface: [WordForm: "unter"]                                 ]

           [   [Allomorph: "lauf"                                     ]]
           [   [             [concatDerivSx: yes]                     ]]
           [   [             [concatInflSx:  yes]                     ]]
           [   [             [concatStem:    yes]                     ]]
           [   [             [Diminutive:    no ]                     ]]
           [   [Combi:       [PlDatSx:       no ]                     ]]
```

Figure 6. Malaga: Sample screen.

The result is a list (indicated by the corner brackets) of analyses. Each analysis is a record (indicated by the square brackets) containing feature-value pairs. In the above example the result list contains exactly one analysis with the following information:

- the analysis type, in this case `Parsed`, i.e. the word form was recognized by the LAG rule mechanism (other possibilities would be `unknown` and `Hypothesis`)
- the segmented surface of the word form
- the part-of-speech tag
- the base form
- a weight which can be used for disambiguation of word forms that have more than one reading; the weight is based on heuristics that evaluate the concatenation processes
- gender of the noun
- case and number of the noun

Malaga is freely available through the GNU general public license and can be found at http://www.linguistik.uni-erlangen.de/~bjoern/Malaga.en.html.
DDM is available at: http://www.linguistik.uni-erlangen.de/~orlorenz/DMM/DMM.en.html.

3.6 Finite state utilities by Jan Daciuk

Jan Daciuk has created a set of programs creating and using finite-state automata for spell-checking, morphological analysis and synthesis, perfect hashing, diacritic restoration, computer-assisted addition of new words into a morphological lexicon etc. Two separate packages, both written in C++, are available: for finite state automata and for transducers (Daciuk, Watson and Watson, 1998). An interface in *elisp* that works with *emacs19* is provided for both of them. Three of these utilities are related to the morphological analysis task:

- **Finita state automata**

  fsa_guess — performs morphological analysis of both known and unknown words. The analysis for known words can be 100% correct, and for the unknown ones — approximative, based on suffixes and prefixes. The dictionary is built from a morphological dictionary for a given language, so once such a dictionary is available, no special linguistic knowledge is required. A Tcl/Tk interface is provided that facilitates the task of adding new words to a dictionary.

  fsa_morph — performs morphological analysis of words, i.e. for a given inflected form, it gives the corresponding lexeme and categories. Several dictionaries may be used at the same time. Dictionaries are compact, data for them has a very simple format (sample *awk* preparation scripts are present in the package), and the program is very fast. *Emacs* interface for text annotation is provided in the package.

- **Transducers**

  tr_morph — performs both morphological analysis and generation of inflected forms. Several dictionaries may be used at the same time. Emacs interface facilitates annotation of corpora.

The utilities can be downloaded via anonymous ftp at ftp://ftp.pg.gda.pl/pub/software/xtras-PG/fsa/. In addition German word list is available from ftp.informatik.tu-muenchen.de:/pub/doc/dict/. The list is 7 bit only, *umlauts* are coded with following *e*, *sharp s* with *ss*. It is difficult to convert them to 8 bit, as not every *oe* is *o umlaut*, not every *ss* is *sharp s*, etc. More information can be found on the Internet at: http://www.pg.gda.pl/~jandac/fsa.html.

3.7 Morphix

The Morphix system was first implemented in 1986 as a programming course by Wolfgang Findler and Günter Neumann. (Neumann and Mazzini, 1999; Neumann et al., 1997; Finkler and Neumann, 1988; Finkler and Neumann, 1986) The system is implemented in Common Lisp and handles all inflectional phenomena of the German language by considering morphologic regularities as the basis for defining fine-grained word-class specific subclassification. It has been tested under Solaris, Linux, Windows 98 and Windows-NT. The German version has a very broad coverage, and an excellent speed (5000 words/sec without compound handling, 2800 words/sec with compound processing (where for each compound all lexically possible decompositions are computed).

Unlike most of the systems considered above that use finite state approach Morphix relies on hierarchical classification and two knowledge sources: *stem lexicon* and *inflectional allomorph* *lexicon* (IAL). The stem lexicon contains classification information for each of the stems known to the system while IAL contains information about the possible morphosyntactic information for a stem given its class. Each entry in the IAL is an *n*-ary tree whose nodes describe the classes and the leaves — the inflectional information. The *n*-ary approach used by Morphix permits the incorporation of both analysis (see Figure 7) and generation in a single system and is claimed to be much faster than the finite-state methods used by the rival systems. The system can be downloaded at http://www.dfki.de/~neumann/morphix/morphix.html.

```
(morph-from-string "Dem Ingenieur ist nichts zu schwoer. ")
```

yields:

```
(((“Dem"
   ("d-det"
    (((:TENSE . :NO) ... (:GENDER . :M) (:NUMBER . :S)
      (:CASE . :DAT))
     ((:TENSE . :NO) ... (:GENDER . :NT)
      (:NUMBER . :S) (:CASE . :DAT)))
    . :DEF))
  ("Ingenieur"
   ("ingenieur"
    (((:TENSE . :NO) ... (:CASE . :NOM))
     ((:TENSE . :NO) ... (:CASE . :DAT))
     ((:TENSE . :NO) ... (:CASE . :AKK)))
    . :N))
  ("ist"
   ("sei"
    (((:TENSE . :PRES) ...
      (:NUMBER . :S) (:CASE . :NO)))
    . :AUX))
  ("nichts" ("nichts" NIL . :PART))
  ("zu" ("zu" NIL . :SUBORD)
   ("zu"
    (((:TENSE . :NO) ... (:CASE . :DAT)))
    . :PREP))
  ("schwoer"
   ("schwoer"
    (((:TENSE . :NO) (:FORM . :IMP) ...
      (:NUMBER . :S) (:CASE . :NO)))
    . :V))
  ("." ("." NIL . :INTP))))
```

Figure 7. Morphix: Morphological Analysis of a sample phrase.

4 The German language

- **Highly inflexional language**

German is a highly inflectional language. In English the nouns change only in number but never according to their case or function in the sentence (except for the possessive case ending *'s*). On the other hand German has 4 cases and each noun changes according to both its case and number. The way a noun changes depends upon both its ending and gender. The notion of gender is irrelevant for most of the nouns in English and it could be determined only for limited noun categories mostly for the living beings.

- **Uniform**

While being a highly inflectional language German is still quite uniform and the noun inflections tend to follow general rules with only few irregularities. Since there are 4 cases and 2 numbers there are up to 8 different forms a noun could theoretically take.

In fact for each German noun some of these forms have the same graphemic representation and thus in general the German nouns have strictly less than 8 different forms. There are some exceptions: Some of the nouns could take up to 4 different forms since they are used in only either singular (e.g. *das Gehren*) or plural (e.g. *die Leute*). Other could have more than 8 forms, which happens when a word can belong to more than one morphological class at the same time. Those are usually words having more than one possible gender with possibly different meaning for the different genders (e.g. *die/der/das Halfter*) although different morphological classes for the same gender are possible as well (e.g. *der Saldo*) mostly for foreign words.

Compared to Bulgarian German is much more uniform. It has only 40 morphological noun classes compared to 72 for Bulgarian while having almost no stem changes except the umlauts and the *ß* alternation which on the other hand happens on a regular principle and a very limited amount of irregular words. In Bulgarian the stem changes are much more common and much more irregular.

- **Easy noun discovery**

According to the German grammar the nouns are always written capitalised. And since this applies only to the nouns and no other part of speech is written capitalised in the general case, the noun discovery is quite simplified compared to languages that do not have this property. We will return to the automatic noun discovery in more details later.

- **Very rich in lexical forms**

Previous multilingual studies have shown German is much richer in different lexical forms than any other Western language. To get an idea of the process we present here a comparison table we borrowed from Adda-Decker and Adda (Adda-Decker M., Adda G., 2000), who collected the data from (Young et al., 1997; Lamel et al., 1995; Matsuoka et al. 1996). Table 2 shows in this particular study German generates 3 times more out-of-vocabulary words than French and 12 times more than English.

	German	**English**	**Italian**	**French**	**Japanese**
Corpus	*FR*	*WSJ*	*Sole 24*	*Le Monde*	*Nikkei*
#words	36M	37.2M	25.7M	37.7M	180M
#distinct	650k	165k	200k	280k	623k
5k cover. %	82.9	90.6	88.3	85.2	88.0
20k cover. %	90.0	97.5	96.3	94.7	96.2
65k cover. %	95.1	99.6	99.0	98.3	99.2
65k-OOV %	4.9	0.4	1.0	1.7	0.8

Table 2. Comparison of 5 languages *(Frankfurter~Rundschau* with *WSJ, Il Sole 24 Ore*, *Le Monde* and *Nikkei* text corpora) in terms of number of distinct words and lexical coverage of the text data for different lexicon sizes. OOV (Out Of Vocabulary) rates are shown for 65k lexicon.

The lexical richness of German is attributed to three major phenomena: inflection, derivation and compounding, the latter being the most powerful since it is generative and can generate theoretically an infinite number of lexical forms. We will return to the issue of compound words discovery and splitting in more details later.

5 Morphological classes

Our morphological classification follows the one developed under the DB-MAT and DBR-MAT projects. The DB-MAT is a German-Bulgarian Machine Translation (MAT) project based on a new MAT-paradigm where the human user is supported by linguistic as well as by subject information, (v. Hahn & Angelova, 1994, 1996). The DBR-MAT is an extension of the DB-MAT project with a new language: Romanian, (Angelova & Bontcheva, 1996a, 1996b). More information about DB-MAT could be found at http://nats-www.informatik.uni-hamburg.de/~dbrmat/db-mat.html, and for DBR-MAT: http://lml.bas.bg/projects/dbr-mat/.

The DB-MAT morphological classes for nouns follow the classification given in *Bulgarisch-Deutsch Wörterbuch*, (Dietmar and Walter, 1987). The dictionary offers 41 classes. (see Table 3)

5.1 Notation

(") in suffix is a signal for application of one of the rules a->ä, o->ö, u->ü and au->äu.

[..] denotes non-obligatory element.

(..) denotes some additional rules to be applied, the rules are encoded by:

1 – concerns the [e]-information in "gen sg", "masc/neut" and means:

a) when the basic form ends with "s / ß / sch / x / chs / z / tz" then vowel "e" is obligatory.

b) when ß stays after a short vowel in the basic form, it is written as "ss" in all forms of the paradigm.

2 – concerns the suffix in "dat pl", "masc/neut" and means: when the basic form ends with "n" there is no second "n" as "dat pl" suffix.

Remark:

Rule 1a) is not obligatory. This is just a preference. In case we generate a text it is better to respect it. But in case we try to reverse an inflection there is no reason to apply it since both forms are in fact legal in German. ?

Class	Singular				Plural				Example Stem
	nom	gen	dat	akk	nom	gen	dat	akk	
m1	0	[e]s(1)	[e]	0	e	e	en	e	Tag
m1a	0	ses	[se]	0	se	se	sen	se	Bus
m2	0	[e]s(1)	[e]	0	"e	"e	"en	"e	Bach
m3	0	[e]s(1)	[e]	0	"er	"er	"ern	"er	Wald
m3a	0	[e]s(1)	[e]	0	er	er	ern	er	Leib
m4	0	s	0	0	0	0	n(2)	0	Deckel
m5	0	s	0	0	"	"	"n(2)	"	Vater
m6	0	s	0	0	s	s	s	s	Gummi
m7	[r]	n	n	n	n	n	n	n	Bekannte
m7a	0	ns	n	n	n	n	n	n	Gedanke
m8	0	en	en	en	en	en	en	en	Mensch
m9	0	[e]s(1)	[e]	0	en	en	en	en	Staat
~~m9a~~	~~0~~	~~s~~	~~0~~	~~0~~	~~en~~	~~en~~	~~en~~	~~en~~	~~Direktor~~
m10	0	s	0	0	n	n	n	n	Konsul
m11	us	us	us	us	en	en	en	en	Organism
f12	0	0	0	0	e	e	en	e	Drangsal
f13	0	0	0	0	se	se	sen	se	Kenntnis
f14	0	0	0	0	"e	"e	"en	"e	Nacht
f14a	0	0	0	0	"	"	"n	"	Mutter
f15	0	0	0	0	s	s	s	s	Kamera
f15a	a	a	a	a	en	en	en	en	Firm
f16	0	0	0	0	n	n	n	n	Blume
~~f16a~~	~~0~~	~~0~~	~~0~~	~~0~~	~~n~~	~~n~~	~~n~~	~~n~~	~~Energie~~
f17	0	0	0	0	en	en	en	en	Zahl
f18	0	0	0	0	nen	nen	nen	nen	Lehrerin

f19	0	n	n	0	n	n	n	n	Angestellte
n20	0	[e]s(1)	[e]	0	e	e	en	e	Schaf
n20a	0	es	[e]	0	"e	"e	"en	"e	Floß
n21	0	[e]s(1)	[e]	0	er	er	ern	er	Feld
n22	0	[e]s(1)	[e]	0	"er	"er	"ern	"er	Dorf
n23	0	s	0	0	0	0	n(2)	0	Fenster
n23a	0	s	0	0	"	"	"n(2)	"	Kloster
n24	0	s	0	0	s	s	s	s	Auto
n25	0	[e]s(1)	[e]	0	en	en	en	en	Bett
n26	[s]	n	n	0	n	n	n	n	Junge
n27	0	ses	[se]	0	se	se	sen	se	Begräbnis
n28	um	ums	um	um	en	en	en	en	Dat
n28a	a	as	a	a	en	en	en	en	Dram
n29	um	ums	um	um	a	a	a	a	Maxim
n30	0	s	0	0	n	n	n	n	Auge
n31	0	[e]s	0	0	ien	ien	ien	ien	Privileg

Table 3. *D B-MAT* morphological classes, corresponding rules and example stems.

Example

We demonstrate the way these rules are applied taking for example the words *der Tag*, *der Vater, die Firma* and *das Floß*, see Table 4.

stem/class	nom sg	gen sg	dat sg	akk sg	nom pl	gen pl	dat pl	akk pl
Tag/**m1**	Tag	Tags Tages	Tag Tage	Tag	Tage	Tage	Tagen	Tage
Vater/**m5**	Vater	Vaters	Vater	Vater	Väter	Väter	Vätern	Väter
Firm/**f15a**	Firma	Firma	Firma	Firma	Firmen	Firmen	Firmen	Firmen
Floß/**n20a**	Floß	Flosses	Floß Flosse	Floß	Flösse	Flösse	Flössen	Flösse

Table 4. Example: morphological class rules application.

Remark

There are some particularities regarding the rules above. The rules *f16* and *f16a* are absolutely identical and differ only because of the stress: it is on *-ie* in singular and on *-i* in plural. Since we are unable to determine automatically the stress of a word given its graphemic form we decided to conflate the classes *f16* and *f16a*. Thus, for our System both classes live under the common caption *f16*.

A similar situation arises with classes *m9* and *m9a*. They differ because of an optional *e* in both genitive singular and dative singular as well as because of the stress: it moves from the 2nd syllable in singular to the 3rd syllable in plural. Again, since we cannot determine the stress from the graphemic form and the only other difference is on non-obligatory elements, we decided to conflate these two classes under *m9*.

Thus, the System works on 39 instead of the original 41 morphological classes. ?

Each rule is associated a gender denoted by the letters *m*, *f* and *n* in the morphological classes names. There are 14+1 masculine, 10+1 feminine and 15 neuter classes. Each of these three subsets is responsible for the rules of a specific gender. Note that the classes *n24* and *m6* are absolutely identical except the gender.

5.2 The Orthographic Reform

We currently do not take into account the German orthographic reform adopted in 1996. The major changes the reform imposes are:

- *umlauts*

  The umlauts have to be written as a sequence of the corresponding short vowel followed by the letter *e*. Thus, the letters *ä*, *ö*, *ü* have to be replaced by *ae*, *oe* and *ue*.
- *letter ß*

  The letter *ß* is no longer valid and has to be replaced by the sequence *ss*.
- *three consonants rule*

  The reform cancels the *three consonants rule*. Thus, when combining words like *Schiff* and *Fahrt* into a single compound we obtain *Schifffahrt* instead of *Schiffahrt* as used to be according to the old German orthography.

The reform is currently not widely accepted and large parts of the population as well as most of the newspapers keep using the old orthography, which results in the increase of the graphemic variability of German. Adda-Decker&Adda reported they found *Schiffahrt* about 2000 times while *Schifffahrt* occurred about 100 times (Adda-Decker&Adda, 2000). These numbers are obtained from a large 300 M words corpora including: *Deutsche Presse Agentur* (30 M words), *Frankfurter Rundschau* (35 M words) and Berliner Tageszeitung (150 M words) as well as several other texts obtained from the Web.

6 Resources used

6.1 Morphologically annotated corpus

negr@ corpus

The NEGRA corpus is the first German linguistically analysed corpus and consists of approximately 176,000 tokens (10,000 sentences) of German newspaper text, taken from the *Frankfurter Rundschau* as contained in the *CD "Multilingual Corpus 1"* of the *European Corpus Initiative* (http://www.coli.uni-sb.de/sfb378/negra-corpus/cd-info-e.html). It is based on approximately 60,000 tokens that were POS tagged at the *Institut für maschinelle Sprachverarbeitung*, Stuttgart. The corpus was extended, tagged with part-of-speech and completely annotated with syntactic structures. The linguistic analysis of the corpus was generated semi-automatically using techniques developed within the NEGRA project (http://www.coli.uni-sb.de/sfb378/projects/NEGRA-en.html). They are part of a boot-strapping process, enabling the research on automatic learning, the development of robust statistical parsing techniques, and models of human language use, in the SFB and many other projects. The corpus was created in the projects NEGRA (*DFG Sonderforschungsbereich 378, Projekt C3*) and LINC (*Universität des Saarlandes*) in Saarbrücken.

The following different types of information are coded in the corpus:

- **Part-of-Speech Tags.** Uses *Stuttgart-Tubingen-Tagset* (STTS), http://www.coli.uni-sb.de/sfb378/negra-corpus/stts.asc
- **Morphological analysis** (only for the first 60,000 tokens). Uses the *expanded STTS*, http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html
- **Grammatical function in the directly dominating phrase.**
- **Category of non-terminal nodes.**

The corpus is freely available for scientific usage. It is internally stored in a SQL database but is distributed in two essential formats: *export* format and *Penn Treebank* format. Figure 8 demonstrates the way the first three sentences are exported using the export format. More information can be found on the Internet at http://www.coli.uni-sb.de/sfb378/negra-corpus/negra-corpus.html.

```
#BOS 1 15 892541360 1
Mögen                VMFIN   3.Pl.Pres.Konj  HD      508
Puristen             NN      Masc.Nom.Pl.*   NK      505
aller                PIDAT   *.Gen.Pl        NK      500
Musikbereiche        NN      Masc.Gen.Pl.*   NK      500
auch                 ADV     --              MO      508
die                  ART     Def.Fem.Akk.Sg  NK      501
Nase                 NN      Fem.Akk.Sg.*    NK      501
rümpfen                      VVINF   --              HD      506
,                    $,      --              --      0
die                  ART     Def.Fem.Nom.Sg  NK      507
Zukunft                      NN      Fem.Nom.Sg.*    NK      507
der                  ART     Def.Fem.Gen.Sg  NK      502
Musik                NN      Fem.Gen.Sg.*    NK      502
liegt                VVFIN   3.Sg.Pres.Ind   HD      509
für                  APPR    Akk             AC      503
viele                PIDAT   *.Akk.Pl        NK      503
junge                ADJA    Pos.*.Akk.Pl.St NK      503
Komponisten          NN      Masc.Akk.Pl.*   NK      503
im                   APPRART Dat.Masc        AC      504
Crossover-Stil               NN      Masc.Dat.Sg.*   NK      504
.                    $.      --              --      0
#500                 NP      --              GR      505
#501                 NP      --              OA      506
#502                 NP      --              GR      507
#503                 PP      --              MO      509
#504                 PP      --              MO      509
#505                 NP      --              SB      508
#506                 VP      --              OC      508
#507                 NP      --              SB      509
#508                 S       --              MO      509
#509                 S       --              --      0
#EOS 1
#BOS 2 2 899973978 1
Sie                  PPER    3.Pl.*.Nom      SB      504
gehen                VVFIN   3.Pl.Pres.Ind   HD      504
gewagte                      ADJA    Pos.*.Akk.Pl.St NK      500
Verbindungen         NN      Fem.Akk.Pl.*    NK      500
und                  KON     --              CD      502
Risiken                      NN      Neut.Akk.Pl.*   CJ      502
ein                  PTKVZ   --              SVP     504
,                    $,      --              --      0
versuchen            VVFIN   3.Pl.Pres.Ind   HD      505
ihre                 PPOSAT  *.Akk.Pl        NK      501
Möglichkeiten        NN      Fem.Akk.Pl.*    NK      501
auszureizen          VVIZU   --              HD      503
.                    $.      --              --      0
#500                 NP      --              CJ      502
#501                 NP      --              OA      503
#502                 CNP     --              OA      504
#503                 VP      --              OC      505
#504                 S       --              CJ      506
#505                 S       --              CJ      506
#506                 CS      --              --      0
#EOS 2
#BOS 3 3 867229898 1
Folklore             NN      Fem.Akk.Sg.*    CJ      500
,                    $,      --              --      0
Rock                 NN      Masc.Akk.Sg.*   CJ      500
,                    $,      --              --      0
Klassik                      NN      Fem.Akk.Sg.*    CJ      500
und                  KON     --              CD      500
Jazz                 NN      Masc.Akk.Sg.*   CJ      500
zu                   PTKZU   --              PM      501
```

```
vermischen           VVINF   --              HD      501
reicht                       VVFIN   3.Sg.Pres.Ind   HD      507
ihnen                PPER    3.Pl.*.Dat      DA      507
nicht                PTKNEG  --              NG      507
,                    $,      --              --      0
sie                  PPER    3.Pl.*.Nom      SB      505
nutzen                       VVFIN   3.Pl.Pres.Ind   HD      505
die                  ART     Def.Fem.Akk.Sg  NK      502
Elektronik           NN      Fem.Akk.Sg.*    NK      502
und                  KON     --              CD      511
sind                 VAFIN   3.Pl.Pres.Ind   HD      510
sogar                ADV     --              MO      509
dazu                 PROAV   --              PH      508
übergegangen         VVPP    --              HD      509
,                    $,      --              --      0
Instrumente          NN      Neut.Akk.Pl.*   OA      506
selbst                       ADV     --              MO      506
zu                   PTKZU   --              PM      503
bauen                VVINF   --              HD      503
.                    $.      --              --      0
#500                 CNP     --              OA      504
#501                 VZ      --              HD      504
#502                 NP      --              OA      505
#503                 VZ      --              HD      506
#504                 VP      --              SB      507
#505                 S       --              CJ      511
#506                 VP      --              RE      508
#507                 S       --              CJ      511
#508                 PP      --              MO      509
#509                 VP      --              OC      510
#510                 S       --              CJ      511
#511                 CS      --              --      0
#EOS 3
```

Figure 8. NEGRA corpus export format example: three sample sentences.

6.2 Lexicons

6.2.1 Word Lexicon

We assume the Word Lexicon contains a complete list of the closed-class words such as:

- article;
- interjection;
- conjunction;
- pronoun;
- preposition;
- numerical, etc.

In addition the Word Lexicon contains some open-class words such as:

- 1st participle;
- 2nd participle;
- adjective;
- adverb;
- noun;
- verb, etc.

Each line in the Word Lexicon represents one word entry and has the following format:

<word> <POS$_1$> <POS$_2$> ... <POS$_n$> @

where

<word> — a word (may be inflected)
<POS$_i$> — a list of all its corresponding Part-of-Speech tags (see below), i=1,2,...,n
@ — end of entry marker to facilitate the parsing

```
......................
Dittrich EIG @
Dittrichs EIG @
Diva SUB @
Divas SUB @
......................
Docht SUB @
Dochte SUB @
Dochten SUB @
Dochts SUB @
Dock SUB @
Docken SUB @
Dockende SUB @
Dockenden SUB @
Dockens SUB @
Docks SUB @
Documenta SUB @
Dodekaeder SUB @
Dodekaedern SUB @
......................
```

```
......................
do ABK @
doch ADV KON @
docke VER @
docken VER @
dockend PA1 VER @
dockende PA1 @
dockendem PA1 @
dockenden PA1 @
dockender PA1 @
dockendes PA1 @
dockest VER @
docket VER @
dockst VER @
dockt VER @
dockte VER @
dockten VER @
docktest VER @
docktet VER @
......................
```

Figure 9. Word Lexicon (extract).

Table 5 contains the abbreviations for the POS tags used in the annotations in the lexicon. In fact we currently adopted the POS tags used by Lezius and described in (Lezius et al., 1998), since we use its lexicon as base.

Lexicon Tag Abbreviation	Tag Description	Entry Count
ABK	abbreviation	53
ADJ	adjective	64008
ADV	adverb	478
ART	article	12
EIG	proper name	2658
INJ	interjection	18
KON	conjunction	69
NEG	negation	6
PA1	1st participle	75905
PA2	2nd participle	74211
PRO	pronoun	174
PRP	preposition	112
SUB	noun	131154
VER	verb	102996
ZAL	numerical	32
ZUS	verb supplement	152

Table 5. POS tags, their description and entry count in the current lexicon.

We keep the words in the lexicon with the first letter capitalised or non-capitalised depending on how it is likely to be normally written according to its part of speech. Thus, the nouns (SUB) and proper nouns (EIG) are capitalised while the other POS are not. This means that the same word can be present twice in the Word Lexicon: once capitalised (e.g. *Docken SUB*) and once — non-capitalised (*docken VER*). In fact we do not currently exploit this lexicon property but we have some considerations for its future usage and prefer to keep this distinction for the moment.

Usually, the lexicon contains all inflected forms of a word but this is only a recommendation and is not strictly necessary. Of course, the bigger the lexicon the better the results expected. The Word Lexicon is used for several different purposes including:

- **nouns identification**

  Although according to the German grammar all the nouns are always written capitalised not all capitalised words can be considered nouns. Each word in the beginning of a sentence is always written capitalised *regardless* of its POS. On the other hand not all words in a non-starting position in a sentence can be considered as incontestable nouns. (In the part of phrase *"Forum Neue Musik fest" Neue* is capitalised although it is an adjective.) Thus, it is a good idea to check a word against the Word Lexicon first and just then apply heuristics exploiting capitalisation. On the other hand this means that the lexicon must be as complete as possible.

- **compound words splitting**

  The German compound word can be made of a sequence of words from a limited POS: noun, adjective, verb, participle and preposition. When we try to split a compound word we have to check whether the words it is made of are present in the lexicon and if so whether their POS is appropriate.

6.2.2 Stem Lexicon

The Stem Lexicon contains a list of the known stems together with their morphological class. Each entry is printed on a single line and starts with a capitalised stem followed by its morphological class. The stem is separated from its morphological class by tabulation. If a stem has more than one morphological class it will appear in more than one entry: each time with different morphological class. The different morphological classes, if more than one, are separated one from the other by single space character. All the stems in the Stem Lexicon are kept capitalised and the class names are kept non-capitalised although this is not strictly necessary and is not supposed to be exploited in any manner. The Stem Lexicon currently contains 13,147 different entries (13,072 stems, because some stems have more than one morphological class, see below) most of which have a single morphological class. The entries are not sorted in any manner.

```
...................
Organism    m11
Teilnehmer  m4
Schreibung  f17
Mazedonier  m4
Photodiode  f16
Peripherie  f16
Operatorin  f18
Halfter             m4
Halfter             f16
Halfter             n23
Reformator  m9
Reorganisation      f17
Monopolist  m8
Frühzündung f17
Judikation  f17
Karosserie  f16
Konfektion  f17
Landschaft  f17
Präposition f17
Tachometer  n23a
Präexistenz f17
```

```
Missetäter  m4
Plattitüde  f16
Reibfläche  f16
Lokomotive  f16
Proportionalsteuer f16
Schonfrist  f17
Penicillin  n20
Seckbacher  m4
Mortadella  f15
Bohrmaschine        f16
Junghering  m1
Gedächtnis  n27
Menschheit  f17
Mikrogramm  m1
Paragenese  f16
Fehlsichtigkeit     f17
Mitteilung  f17
Seltenheit  f17
Diktiergerät        n20
Nachgebühr  f17
Taufschein  m1
Kleckserei  f17
```

```
Konditorei  f17
Konfession  f17
Endbuchstabe       m7
Kaltleiter  m4
Gegenschrift       f17
Kleptomane  m7
Schwammerl  n23
Hohlleiste  f16
Ministerin  f18
Karfreitag  m1
Intubation  f17
Reparation  f17
Debütantin  f18
Photozelle  f16
Hypotenuse  f16
Buddelschiff       n20
Entführung  f17
Bruttoregistertonne       f16
Feministin  f18
Zusammenstellung   f17
Sicherheit  f17
Auswärtige  m7
Auswärtige  f16
Polarlicht  n21
Rechenautomat      m8
Auszählung  f17
Aufbesserung       f17
Etymologie  f16
Medianwert  m1
Schelmerei  f17
Kaffeesieder       m4
Expedition  f17
Sendereihe  f16
Illuminierung      f17
Enthüllung  f17
Schwindler  m4
Thermalbad  n22
Bezugsquelle       f16
Normalzeit  f17
Erstkommunion      f17
Saxofonist  m8
Respirator  m9
Präsidentin f18
Marionette  f16
Plazierung  f17
Aussöhnung  f17
Rekonstruktion     f17
Kennziffer  f16
Parksünder  m4
Kraftmeier  m4
Nachahmung  f17
Pappbecher  m4
Gefährdung  f17
Hängegleiter       m4
Nebenklage  f16
Innentemperatur    f17
Isolierung  f17
Radfahrweg  m1
Gemeinschaft       f17
Nachgeburt  f17
Direktübertragung  f17
Nettigkeit  f17
Glücksbringer      m4
Klapptisch  m1
Silberlöwe  m7
Kleinstadt  f14
Sammelband  m3
Walldorfer  m4
Bittermittel       n23
Saarländer  m4
Millimeter  m4
Schusterei  f17
Annäherung  f17
Kronprinzessin     f18
Feudalsystem       n20
Resorption  f17
Nebentisch  m1
Seligsprechung     f17
Telefonist  m8
Okkupati/on f17
Interimsregierung  f17
Teilerfolg  m1
Brennelement       n20
...................
```

Figure 10. Stem Lexicon (extract).

Anyway, some of the stems could have more than one morphological class. This is usually due to possibilities for different gender. The current Stem Lexicon contains 74 words having more than one morphological class: *der/die/das Halfter* having three different morphological classes and 73 other stems having two different morphological classes.

```
..................
Halfter m4 f16 n23
Ami m6 f15
Tor m8 n20
Bund m1 n20
Flur m1 f12
Teil m1 n20
Bambi m6 n24
Gelee m6 n24
Bravo f15 n24
Radar m6 n24
Gummi m6 n24
Junge m7 n26
Biotop m1 n20
Filter m4 n23
Bummel m4 f16
Fremde m7 f16
Laster m4 n23
Messer m4 n23
Moment m1 n20
```

```
Kalkül m1 n20
Single m6 f15
Steuer f16 n23
Trikot m6 n24
Knäuel m4 n23
Tüpfel m4 n23
Elf m8 f17
Gig f15 n24
Keks m1 n20
Tote m7 f16
Cartoon m6 n24
Break m6 n24
Lauch m1 n20
Liter m4 n23
Weise m7 f16
Dropout m6 n24
Joghurt m6 n24
Rebhuhn m3 n22
Krempel m4 f16
Praktik f17 n29
Torpedo m6 n24
Bonbon m6 n24
Mangel m5 f16
Dotter m4 n23
Poster m4 n23
Beigeordnete m7 f16
Heide m7 f16
Sakko m6 n24
Bekannte m7 f16
Geliebte m7 f16
Deutsche m7 f19
Leiter m4 f16
Raster m4 n23
Urahne m7 f16
Kristall m1 n20
Farbige m7 f16
Erbteil m1 n20
Kunde m7 f16
Taube m7 f16
Katapult m1 n20
Verdienst m1 n20
Verhau m1 n20
Abgeordnete m7 f16
Angestellte m7 f16
Techtelmechtel m4 n23
Schlamassel m4 n23
Behinderte m7 f16
Delegierte m7 f16
Angeklagte m7 f16
Vorsitzende m7 f16
Angehörige m7 f16
Obdachlose m7 f16
Auswärtige m7 f16
Verbündete m7 f16
Sachverständige m7 f16
...................
```

Figure 11. Stems having more than one morphological class *(compressed in one line to save space)*

It is also possible that a noun has several different morphological classes all being from the same gender. An example is the foreign word *der Saldo*. The problem is with its plural ending. It could take the ending *-s*, which implies the morphological class *m6*. But it could also take *-en*, which leads to *m9*. And finally, it could also take the foreign ending *-i*, which cannot be covered by any of the German morphological classes. (see Table 6)

class	**nom sg**	**gen sg**	**dat sg**	**akk sg**	**nom pl**	**gen pl**	**dat pl**	**akk pl**
m6 *Saldo*	Saldo	Saldos	Saldo	Saldo	Saldos	Saldos	Saldos	Saldos
m9 *Saldo*	Saldo	Saldos	Saldo	Saldo	Salden	Salden	Salden	Salden
?? *Saldo*	Saldo	Saldos	Saldo	Saldo	Saldi	Saldi	Saldi	Saldi

Table 6. The morphological classes for *der Saldo*.

The current Stem Lexicon has been induced automatically in a way that does not permit for a stem to have more than one morphological class having the same gender. Thus, *der Saldo* is present in our lexicon only with the morphological class *m6*. We will return to this issue below.

6.2.3 Expanded Stem Lexicon

The Expanded Stem Lexicon is an expansion of the Stem Lexicon. This is a generated list of all the forms of the word's declination. Usually, they are 8, one form per case/number combination, but sometimes could be 9 or 10 since some of the rules have optional elements (especially in gen/sg). The classes *m1a*, *m7*, *n20a*, *n26*, *n27*, *n31* have one optional element and thus 9 forms, and *m1*, *m2*, *m3*, *m3a*, *m9*, *n20*, *n21*, *n22*, *n25* have 10 forms. Each word has a corresponding record in the lexicon, usually 10 lines long (but sometimes 11 or 12). The general format is shown in Figure 12.

```
@
<basic_form>
<nom_sg_form>  NOM    SIN    <gender>
<gen_sg_form>  GEN    SIN    <gender>
<dat_sg_form>  DAT    SIN    <gender>
<akk_sg_form>  AKK    SIN    <gender>
<nom_pl_form>  NOM    PLU    <gender>
<gen_pl_form>  GEN    PLU    <gender>
<dat_pl_form>  DAT    PLU    <gender>
<akk_pl_form>  AKK    PLU    <gender>
```

Figure 12. General Expanded Stem Lexicon format.

Figure 13 shows the entry for *der Organismus*.

```
@
Organismus
Organismus       NOM      SIN      MAS
Organismus       GEN      SIN      MAS
Organismus       DAT      SIN      MAS
Organismus       AKK      SIN      MAS
Organismen       NOM      PLU      MAS
Organismen       GEN      PLU      MAS
Organismen       DAT      PLU      MAS
Organismen       AKK      PLU      MAS
```

Figure 13. Expanded Stem Lexicon, *der Organismus*.

The entries in the Expanded Stem Lexicon are not sorted in any manner but appear in the same sequence as the entries in the Stem Lexicon do. This allows an easy identification of the correspondence between the stems and their expansions. In fact this is not strictly necessary since the stem and the base form differ in only few cases (see Table 3) and can be obtained automatically one from the other. If we know the stem and its class we know how to obtain the nom/sg form, which is the base form. On the other hand if we know the base form and all its inflections we can obtain its morphological class and from there decide whether we have to cut something from the base form. Anyway, this is not so straightforward and we decided to impose the same order on both lexicons. In fact we could just output the stem instead of the base form but we are willing to keep the Expanded Stem Lexicon as human readable as possible since it is generated automatically and we would like to be able to easily check its contents. Thus, words like *der Organismus* are listed as *Organismus* and not as *Organism*. The words that have more than one morphological class appear once per each class. Thus, *der/die/das Halfter* appears 3 times one after the other.

The lexicon in its present format is unnecessarily huge. What we really need is just a list of all distinct word forms that could be generated given an entry from the Stem Lexicon, which is a stem and list of its possible morphological classes. Thus, a much more compact entry form could be used (no more need for the separator `@` as well):

```
<stem> <word_form_1> <word_form_2> ... <word_form_n>
```

Thus, the entry for *der Organismus* would be now represented by:

```
Organism Organismus Organismen
```

What is really important is that the Expanded Stem Lexicon must list *all* the forms whose stems are known. The same applies to the Word Lexicon: all the words from Expanded Stem Lexicon must be included in the Word Lexicon. We rely on these properties to reject the known stems as candidates for the unknown words: An *unknown* word cannot have a known stem since all the words that have this stem are supposed to be included in both the Expanded Stem Lexicon and the Word Lexicon and thus are *known*. This means the stem is not appropriate for the word in question and has to be rejected (in fact it is not appropriate for any unknown word). We will return to this issue in more details later.

```
@
Verschiebung
Verschiebung    NOM     SIN     FEM
Verschiebung    GEN     SIN     FEM
Verschiebung    DAT     SIN     FEM
Verschiebung    AKK     SIN     FEM
Verschiebungen  NOM     PLU     FEM
Verschiebungen  GEN     PLU     FEM
Verschiebungen  DAT     PLU     FEM
Verschiebungen  AKK     PLU     FEM
@
Knappschaftskasse
Knappschaftskasse       NOM     SIN
        FEM
Knappschaftskasse       GEN     SIN
        FEM
Knappschaftskasse       DAT     SIN
        FEM
Knappschaftskasse       AKK     SIN
        FEM
Knappschaftskassen      NOM     PLU
        FEM
Knappschaftskassen      GEN     PLU
        FEM
Knappschaftskassen      DAT     PLU
        FEM
Knappschaftskassen      AKK     PLU
        FEM
@
A
A       NOM     SIN     NEU
As      GEN     SIN     NEU
A       DAT     SIN     NEU
A       AKK     SIN     NEU
As      NOM     PLU     NEU
As      GEN     PLU     NEU
As      DAT     PLU     NEU
As      AKK     PLU     NEU
@
Weltenbummler
Weltenbummler   NOM     SIN     MAS
Weltenbummlers  GEN     SIN     MAS
Weltenbummler   DAT     SIN     MAS
Weltenbummler   AKK     SIN     MAS
Weltenbummler   NOM     PLU     MAS
Weltenbummler   GEN     PLU     MAS
Weltenbummlern  DAT     PLU     MAS
Weltenbummler   AKK     PLU     MAS
@
Verschiffung
Verschiffung    NOM     SIN     FEM
Verschiffung    GEN     SIN     FEM
Verschiffung    DAT     SIN     FEM
Verschiffung    AKK     SIN     FEM
Verschiffungen  NOM     PLU     FEM
Verschiffungen  GEN     PLU     FEM
Verschiffungen  DAT     PLU     FEM
Verschiffungen  AKK     PLU     FEM
@
O
```

```
O       NOM     SIN     NEU
Os      GEN     SIN     NEU
O       DAT     SIN     NEU
O       AKK     SIN     NEU
Os      NOM     PLU     NEU
Os      GEN     PLU     NEU
Os      DAT     PLU     NEU
Os      AKK     PLU     NEU
@
Zeichenblock
Zeichenblock    NOM     SIN     MAS
Zeichenblocks   GEN     SIN     MAS
Zeichenblocke   DAT     SIN     MAS
Zeichenblock    DAT     SIN     MAS
Zeichenblock    AKK     SIN     MAS
Zeichenblöcke   NOM     PLU     MAS
Zeichenblöcke   GEN     PLU     MAS
Zeichenblöcken  DAT     PLU     MAS
Zeichenblöcke   AKK     PLU     MAS
@
Fernlastfahrer
Fernlastfahrer  NOM     SIN     MAS
Fernlastfahrers GEN     SIN     MAS
Fernlastfahrer  DAT     SIN     MAS
Fernlastfahrer  AKK     SIN     MAS
Fernlastfahrer  NOM     PLU     MAS
Fernlastfahrer  GEN     PLU     MAS
Fernlastfahrern DAT     PLU     MAS
Fernlastfahrer  AKK     PLU     MAS
@
AG
AG      NOM     SIN     FEM
AG      GEN     SIN     FEM
AG      DAT     SIN     FEM
AG      AKK     SIN     FEM
AGs     NOM     PLU     FEM
AGs     GEN     PLU     FEM
AGs     DAT     PLU     FEM
AGs     AKK     PLU     FEM
@
CD
CD      NOM     SIN     FEM
CD      GEN     SIN     FEM
CD      DAT     SIN     FEM
CD      AKK     SIN     FEM
CDs     NOM     PLU     FEM
CDs     GEN     PLU     FEM
CDs     DAT     PLU     FEM
CDs     AKK     PLU     FEM
@
Organismus
Organismus      NOM     SIN     MAS
Organismus      GEN     SIN     MAS
Organismus      DAT     SIN     MAS
Organismus      AKK     SIN     MAS
Organismen      NOM     PLU     MAS
Organismen      GEN     PLU     MAS
Organismen      DAT     PLU     MAS
Organismen      AKK     PLU     MAS
@
```

```
Teilnehmer
Teilnehmer       NOM     SIN     MAS
Teilnehmers      GEN     SIN     MAS
Teilnehmer       DAT     SIN     MAS
Teilnehmer       AKK     SIN     MAS
Teilnehmer       NOM     PLU     MAS
Teilnehmer       GEN     PLU     MAS
Teilnehmern      DAT     PLU     MAS
Teilnehmer       AKK     PLU     MAS
@
Halfter
Halfter NOM      SIN     MAS
Halfters         GEN     SIN     MAS
Halfter DAT      SIN     MAS
Halfter AKK      SIN     MAS
Halfter NOM      PLU     MAS
Halfter GEN      PLU     MAS
Halftern         DAT     PLU     MAS
Halfter AKK      PLU     MAS
@
```

```
Halfter
Halfter NOM      SIN     FEM
Halfter GEN      SIN     FEM
Halfter DAT      SIN     FEM
Halfter AKK      SIN     FEM
Halftern         NOM     PLU     FEM
Halftern         GEN     PLU     FEM
Halftern         DAT     PLU     FEM
Halftern         AKK     PLU     FEM
@
Halfter
Halfter NOM      SIN     NEU
Halfters         GEN     SIN     NEU
Halfter DAT      SIN     NEU
Halfter AKK      SIN     NEU
Halfter NOM      PLU     NEU
Halfter GEN      PLU     NEU
Halftern         DAT     PLU     NEU
Halfter AKK      PLU     NEU
@
```

Figure 14. Expanded Stem Lexicon (extract).

7 Lexicons creation

7.1 Morphy Lexicon

We used the free lexicon of the morphological system Morphy by Lezius, which contains 50,597 stems (17380 nouns, 22184 adjectives, 1409 proper nouns etc.) and 324,000 different word forms. (*In fact Morphy gave us 24975 nouns + proper nouns stems when asked to extract its lexicon, which means the numbers reported above are lower than the reality*.) The lexicon is stored in compressed form as stem its part of speech and morphological rules saying how to generate the coresponding word forms. Neither the morphological generation rules nor the file format are described in the Lezius papers but the system offers the option to export (parts of) the lexicon. One has as well the option to choose between the large set (about 1000 tags) and the small set (51 tags). We used the large Morphy tag set to generate our lexicons: Word Lexicon, Stem Lexicon and Expanded Stem Lexicon.

The creation of the Word Lexicon in the format described above is quite easy but is not so straightforward as one may initially think. To get an idea of the output Morphy generates when asked to export its whole lexicon we give here few examples on Figure 15 (the whole output file size is 292 MB!). Since we are interested in a limited subset of the tags as given in Table 5 we designed a simple program to clean the unnessessary noise.

```
@
Weltenbummler
Weltenbummler SUB NOM SIN MAS
Weltenbummler SUB DAT SIN MAS
Weltenbummler SUB AKK SIN MAS
Weltenbummler SUB NOM PLU MAS
Weltenbummler SUB GEN PLU MAS
Weltenbummler SUB AKK PLU MAS
@
Zusammenreimenden
*zusammenreimend PA1 GEN SIN MAS GRU SOL VER (zusammen)reimen
*zusammenreimend PA1 GEN SIN NEU GRU SOL VER (zusammen)reimen
*zusammenreimend PA1 AKK SIN MAS GRU SOL VER (zusammen)reimen
*zusammenreimend PA1 DAT PLU MAS GRU SOL VER (zusammen)reimen
*zusammenreimend PA1 DAT PLU FEM GRU SOL VER (zusammen)reimen
*zusammenreimend PA1 DAT PLU NEU GRU SOL VER (zusammen)reimen
```

```
*zusammenreimend PA1 GEN SIN MAS GRU DEF VER (zusammen)reimen
Zusammenreimende SUB GEN SIN MAS ADJ zusammenreimend
*zusammenreimend PA1 GEN SIN FEM GRU DEF VER (zusammen)reimen
Zusammenreimende SUB GEN SIN FEM ADJ zusammenreimend
*zusammenreimend PA1 GEN SIN NEU GRU DEF VER (zusammen)reimen
Zusammenreimende SUB GEN SIN NEU ADJ zusammenreimend
*zusammenreimend PA1 DAT SIN MAS GRU DEF VER (zusammen)reimen
Zusammenreimende SUB DAT SIN MAS ADJ zusammenreimend
*zusammenreimend PA1 DAT SIN FEM GRU DEF VER (zusammen)reimen
Zusammenreimende SUB DAT SIN FEM ADJ zusammenreimend
*zusammenreimend PA1 DAT SIN NEU GRU DEF VER (zusammen)reimen
Zusammenreimende SUB DAT SIN NEU ADJ zusammenreimend
*zusammenreimend PA1 AKK SIN MAS GRU DEF VER (zusammen)reimen
Zusammenreimende SUB AKK SIN MAS ADJ zusammenreimend
*zusammenreimend PA1 NOM PLU MAS GRU DEF VER (zusammen)reimen
Zusammenreimende SUB NOM PLU MAS ADJ zusammenreimend
*zusammenreimend PA1 NOM PLU FEM GRU DEF VER (zusammen)reimen
Zusammenreimende SUB NOM PLU FEM ADJ zusammenreimend
*zusammenreimend PA1 NOM PLU NEU GRU DEF VER (zusammen)reimen
Zusammenreimende SUB NOM PLU NEU ADJ zusammenreimend
*zusammenreimend PA1 GEN PLU MAS GRU DEF VER (zusammen)reimen
Zusammenreimende SUB GEN PLU MAS ADJ zusammenreimend
*zusammenreimend PA1 GEN PLU FEM GRU DEF VER (zusammen)reimen
Zusammenreimende SUB GEN PLU FEM ADJ zusammenreimend
*zusammenreimend PA1 GEN PLU NEU GRU DEF VER (zusammen)reimen
Zusammenreimende SUB GEN PLU NEU ADJ zusammenreimend
*zusammenreimend PA1 DAT PLU MAS GRU DEF VER (zusammen)reimen
Zusammenreimende SUB DAT PLU MAS ADJ zusammenreimend
*zusammenreimend PA1 DAT PLU FEM GRU DEF VER (zusammen)reimen
Zusammenreimende SUB DAT PLU FEM ADJ zusammenreimend
*zusammenreimend PA1 DAT PLU NEU GRU DEF VER (zusammen)reimen
Zusammenreimende SUB DAT PLU NEU ADJ zusammenreimend
*zusammenreimend PA1 AKK PLU MAS GRU DEF VER (zusammen)reimen
Zusammenreimende SUB AKK PLU MAS ADJ zusammenreimend
*zusammenreimend PA1 AKK PLU FEM GRU DEF VER (zusammen)reimen
Zusammenreimende SUB AKK PLU FEM ADJ zusammenreimend
*zusammenreimend PA1 AKK PLU NEU GRU DEF VER (zusammen)reimen
Zusammenreimende SUB AKK PLU NEU ADJ zusammenreimend
*zusammenreimend PA1 GEN SIN MAS GRU IND VER (zusammen)reimen
*zusammenreimend PA1 GEN SIN FEM GRU IND VER (zusammen)reimen
*zusammenreimend PA1 GEN SIN NEU GRU IND VER (zusammen)reimen
*zusammenreimend PA1 DAT SIN MAS GRU IND VER (zusammen)reimen
*zusammenreimend PA1 DAT SIN FEM GRU IND VER (zusammen)reimen
*zusammenreimend PA1 DAT SIN NEU GRU IND VER (zusammen)reimen
*zusammenreimend PA1 AKK SIN MAS GRU IND VER (zusammen)reimen
*zusammenreimend PA1 NOM PLU MAS GRU IND VER (zusammen)reimen
*zusammenreimend PA1 NOM PLU FEM GRU IND VER (zusammen)reimen
*zusammenreimend PA1 NOM PLU NEU GRU IND VER (zusammen)reimen
*zusammenreimend PA1 GEN PLU MAS GRU IND VER (zusammen)reimen
*zusammenreimend PA1 GEN PLU FEM GRU IND VER (zusammen)reimen
*zusammenreimend PA1 GEN PLU NEU GRU IND VER (zusammen)reimen
*zusammenreimend PA1 DAT PLU MAS GRU IND VER (zusammen)reimen
*zusammenreimend PA1 DAT PLU FEM GRU IND VER (zusammen)reimen
*zusammenreimend PA1 DAT PLU NEU GRU IND VER (zusammen)reimen
*zusammenreimend PA1 AKK PLU MAS GRU IND VER (zusammen)reimen
*zusammenreimend PA1 AKK PLU FEM GRU IND VER (zusammen)reimen
*zusammenreimend PA1 AKK PLU NEU GRU IND VER (zusammen)reimen
@
Verkappten
*verkappen VER 1 PLU PRT SFT
*verkappen VER 1 PLU KJ2 SFT
*verkappen VER 3 PLU PRT SFT
*verkappen VER 3 PLU KJ2 SFT
*verkappt PA2 GEN SIN MAS GRU SOL VER verkappen
*verkappt PA2 GEN SIN NEU GRU SOL VER verkappen
*verkappt PA2 AKK SIN MAS GRU SOL VER verkappen
*verkappt PA2 DAT PLU MAS GRU SOL VER verkappen
*verkappt PA2 DAT PLU FEM GRU SOL VER verkappen
*verkappt PA2 DAT PLU NEU GRU SOL VER verkappen
*verkappt PA2 GEN SIN MAS GRU DEF VER verkappen
Verkappte SUB GEN SIN MAS ADJ verkappt
```

```
*verkappt PA2 GEN SIN FEM GRU DEF VER verkappen
Verkappte SUB GEN SIN FEM ADJ verkappt
*verkappt PA2 GEN SIN NEU GRU DEF VER verkappen
Verkappte SUB GEN SIN NEU ADJ verkappt
*verkappt PA2 DAT SIN MAS GRU DEF VER verkappen
Verkappte SUB DAT SIN MAS ADJ verkappt
*verkappt PA2 DAT SIN FEM GRU DEF VER verkappen
Verkappte SUB DAT SIN FEM ADJ verkappt
*verkappt PA2 DAT SIN NEU GRU DEF VER verkappen
Verkappte SUB DAT SIN NEU ADJ verkappt
*verkappt PA2 AKK SIN MAS GRU DEF VER verkappen
Verkappte SUB AKK SIN MAS ADJ verkappt
*verkappt PA2 NOM PLU MAS GRU DEF VER verkappen
Verkappte SUB NOM PLU MAS ADJ verkappt
*verkappt PA2 NOM PLU FEM GRU DEF VER verkappen
Verkappte SUB NOM PLU FEM ADJ verkappt
*verkappt PA2 NOM PLU NEU GRU DEF VER verkappen
Verkappte SUB NOM PLU NEU ADJ verkappt
*verkappt PA2 GEN PLU MAS GRU DEF VER verkappen
Verkappte SUB GEN PLU MAS ADJ verkappt
*verkappt PA2 GEN PLU FEM GRU DEF VER verkappen
Verkappte SUB GEN PLU FEM ADJ verkappt
*verkappt PA2 GEN PLU NEU GRU DEF VER verkappen
Verkappte SUB GEN PLU NEU ADJ verkappt
*verkappt PA2 DAT PLU MAS GRU DEF VER verkappen
Verkappte SUB DAT PLU MAS ADJ verkappt
*verkappt PA2 DAT PLU FEM GRU DEF VER verkappen
Verkappte SUB DAT PLU FEM ADJ verkappt
*verkappt PA2 DAT PLU NEU GRU DEF VER verkappen
Verkappte SUB DAT PLU NEU ADJ verkappt
*verkappt PA2 AKK PLU MAS GRU DEF VER verkappen
Verkappte SUB AKK PLU MAS ADJ verkappt
*verkappt PA2 AKK PLU FEM GRU DEF VER verkappen
Verkappte SUB AKK PLU FEM ADJ verkappt
*verkappt PA2 AKK PLU NEU GRU DEF VER verkappen
Verkappte SUB AKK PLU NEU ADJ verkappt
*verkappt PA2 GEN SIN MAS GRU IND VER verkappen
*verkappt PA2 GEN SIN FEM GRU IND VER verkappen
*verkappt PA2 GEN SIN NEU GRU IND VER verkappen
*verkappt PA2 DAT SIN MAS GRU IND VER verkappen
*verkappt PA2 DAT SIN FEM GRU IND VER verkappen
*verkappt PA2 DAT SIN NEU GRU IND VER verkappen
*verkappt PA2 AKK SIN MAS GRU IND VER verkappen
*verkappt PA2 NOM PLU MAS GRU IND VER verkappen
*verkappt PA2 NOM PLU FEM GRU IND VER verkappen
*verkappt PA2 NOM PLU NEU GRU IND VER verkappen
*verkappt PA2 GEN PLU MAS GRU IND VER verkappen
*verkappt PA2 GEN PLU FEM GRU IND VER verkappen
*verkappt PA2 GEN PLU NEU GRU IND VER verkappen
*verkappt PA2 DAT PLU MAS GRU IND VER verkappen
*verkappt PA2 DAT PLU FEM GRU IND VER verkappen
*verkappt PA2 DAT PLU NEU GRU IND VER verkappen
*verkappt PA2 AKK PLU MAS GRU IND VER verkappen
*verkappt PA2 AKK PLU FEM GRU IND VER verkappen
*verkappt PA2 AKK PLU NEU GRU IND VER verkappen
@
Geschiedeneren
*geschieden PA2 GEN SIN MAS KOM SOL VER scheiden
*geschieden PA2 GEN SIN NEU KOM SOL VER scheiden
*geschieden PA2 AKK SIN MAS KOM SOL VER scheiden
*geschieden PA2 DAT PLU MAS KOM SOL VER scheiden
*geschieden PA2 DAT PLU FEM KOM SOL VER scheiden
*geschieden PA2 DAT PLU NEU KOM SOL VER scheiden
*geschieden PA2 GEN SIN MAS KOM DEF VER scheiden
Geschiedenere SUB GEN SIN MAS ADJ geschieden
*geschieden PA2 GEN SIN FEM KOM DEF VER scheiden
Geschiedenere SUB GEN SIN FEM ADJ geschieden
*geschieden PA2 GEN SIN NEU KOM DEF VER scheiden
Geschiedenere SUB GEN SIN NEU ADJ geschieden
*geschieden PA2 DAT SIN MAS KOM DEF VER scheiden
Geschiedenere SUB DAT SIN MAS ADJ geschieden
```

```
*geschieden PA2 DAT SIN FEM KOM DEF VER scheiden
Geschiedenere SUB DAT SIN FEM ADJ geschieden
*geschieden PA2 DAT SIN NEU KOM DEF VER scheiden
Geschiedenere SUB DAT SIN NEU ADJ geschieden
*geschieden PA2 AKK SIN MAS KOM DEF VER scheiden
Geschiedenere SUB AKK SIN MAS ADJ geschieden
*geschieden PA2 NOM PLU MAS KOM DEF VER scheiden
Geschiedenere SUB NOM PLU MAS ADJ geschieden
*geschieden PA2 NOM PLU FEM KOM DEF VER scheiden
Geschiedenere SUB NOM PLU FEM ADJ geschieden
*geschieden PA2 NOM PLU NEU KOM DEF VER scheiden
Geschiedenere SUB NOM PLU NEU ADJ geschieden
*geschieden PA2 GEN PLU MAS KOM DEF VER scheiden
Geschiedenere SUB GEN PLU MAS ADJ geschieden
*geschieden PA2 GEN PLU FEM KOM DEF VER scheiden
Geschiedenere SUB GEN PLU FEM ADJ geschieden
*geschieden PA2 GEN PLU NEU KOM DEF VER scheiden
Geschiedenere SUB GEN PLU NEU ADJ geschieden
*geschieden PA2 DAT PLU MAS KOM DEF VER scheiden
Geschiedenere SUB DAT PLU MAS ADJ geschieden
*geschieden PA2 DAT PLU FEM KOM DEF VER scheiden
Geschiedenere SUB DAT PLU FEM ADJ geschieden
*geschieden PA2 DAT PLU NEU KOM DEF VER scheiden
Geschiedenere SUB DAT PLU NEU ADJ geschieden
*geschieden PA2 AKK PLU MAS KOM DEF VER scheiden
Geschiedenere SUB AKK PLU MAS ADJ geschieden
*geschieden PA2 AKK PLU FEM KOM DEF VER scheiden
Geschiedenere SUB AKK PLU FEM ADJ geschieden
*geschieden PA2 AKK PLU NEU KOM DEF VER scheiden
Geschiedenere SUB AKK PLU NEU ADJ geschieden
*geschieden PA2 GEN SIN MAS KOM IND VER scheiden
*geschieden PA2 GEN SIN FEM KOM IND VER scheiden
*geschieden PA2 GEN SIN NEU KOM IND VER scheiden
*geschieden PA2 DAT SIN MAS KOM IND VER scheiden
*geschieden PA2 DAT SIN FEM KOM IND VER scheiden
*geschieden PA2 DAT SIN NEU KOM IND VER scheiden
*geschieden PA2 AKK SIN MAS KOM IND VER scheiden
*geschieden PA2 NOM PLU MAS KOM IND VER scheiden
*geschieden PA2 NOM PLU FEM KOM IND VER scheiden
*geschieden PA2 NOM PLU NEU KOM IND VER scheiden
*geschieden PA2 GEN PLU MAS KOM IND VER scheiden
*geschieden PA2 GEN PLU FEM KOM IND VER scheiden
*geschieden PA2 GEN PLU NEU KOM IND VER scheiden
*geschieden PA2 DAT PLU MAS KOM IND VER scheiden
*geschieden PA2 DAT PLU FEM KOM IND VER scheiden
*geschieden PA2 DAT PLU NEU KOM IND VER scheiden
*geschieden PA2 AKK PLU MAS KOM IND VER scheiden
*geschieden PA2 AKK PLU FEM KOM IND VER scheiden
*geschieden PA2 AKK PLU NEU KOM IND VER scheiden
```

Figure 15. Morphy lexicon extraction output.

7.2 Automatic Morphological Classes Induction

We built the Stem Lexicon automatically using the Morphy lexicon. For each morphological class successfully induced for a given stem we wrote a corresponding record in the Expanded Stem Lexicon. The morphological classes have been induced automatically from the word forms and their corresponding morphological tags in the Morphy lexicon. For each group of inflected nouns or proper names (the proper nouns in German in general change in case/number and follow the general rules defined by the morphological classes) sharing a common base form we tried to find a corresponding pair of stem and morphological class that could generate these forms.

A morphological class was induced if and only if both conditions below hold:

1) **The word has at least one form for *each* of the 8 possible combinations of number/case the gender being fixed.**

Thus, we want that there is a word form for each of: sg/nom, sg/gen, sg/dat, sg/akk, pl/nom, pl/gen, pl/dat, pl/akk. Since some of the morphological classes contain non-obligatory elements it is possible that there is more than one form for some of these (see Table 4). On the other hand there are some words that are used in either only singular or only plural (see below). We cannot classify unambiguously any of these since they will be covered by a set of classes.

Some of the words could have more than one gender (e.g. *der/die/das Halfter*). The induction strategy used can induce in cases like this a different morphological class for each gender. But we are currently unable, due to forms overlap, to induce more than one morphological class for the same gender, although this phenomenon is possible for words like *der Saldo* (see Table 6). (In fact *der Saldo* is met in the Morphy lexicon with exactly 8 forms all covered by the class m6.)

2) **The stem sel?cted must cover at least one word form for each of the 8 combinations of case and number given the gender.**

This is important since as have been noted above there are words in the Morphy lexicon that could have more then one gender. We try to induce automatically a morphological class for each of the genders separately. In case there are more than one morphological classes for the same gender we will induce only one of them.

Remark

We would like to stress again that we work with only 39 instead of the original 41 DB-MAT morphological classes (see Table 3): *f16* and *f16a* are conflated under *f16*, *m9* and *m9a* — under *m9*.?

Some of the words have more than one gender and thus more than one morphological class. These cases are handled appropriately and in case enough forms are available for some gender the corresponding class has been induced. For example Figure 16 shows the data in the Morphy lexicon (we filtered only the nouns) for the base word form *Halfter*.

```
@
Halfter
Halfter SUB NOM SIN MAS
Halfter SUB DAT SIN MAS
Halfter SUB AKK SIN MAS
Halfter SUB NOM PLU MAS
Halfter SUB GEN PLU MAS
Halfter SUB AKK PLU MAS
Halfter SUB NOM SIN FEM
Halfter SUB GEN SIN FEM
Halfter SUB DAT SIN FEM
Halfter SUB AKK SIN FEM
Halfter SUB NOM SIN NEU
Halfter SUB DAT SIN NEU
Halfter SUB AKK SIN NEU
Halfter SUB NOM PLU NEU
Halfter SUB GEN PLU NEU
Halfter SUB AKK PLU NEU
@
Halftern
Halftern SUB NOM SIN NEU INF
Halftern SUB DAT SIN NEU INF
Halftern SUB AKK SIN NEU INF
Halfter SUB DAT PLU MAS
Halfter SUB NOM PLU FEM
Halfter SUB GEN PLU FEM
Halfter SUB DAT PLU FEM
Halfter SUB AKK PLU FEM
Halfter SUB DAT PLU NEU
@
Halfterns
Halftern SUB GEN SIN NEU INF
```

```
@
Halfters
Halfter SUB GEN SIN MAS
Halfter SUB GEN SIN NEU
```

Figure 16. Entries for Halfter in the Morphy lexicon. Only the nouns have been filtered.

Thus, the classes *m4*, *f16* and *n23* have been assigned to the stem *Halfter* (which is the same as the base form) and the corresponding three lines have been added to the Stem Lexicon, see Figure 17.

```
Halfter      m4
Halfter      f16
Halfter      n23
```

Figure 17. Entry for *Halfter* in the Stem Lexicon.

At the same time a set of lines has been added to the Expanded Stem Lexicon, see Figure 18.

```
@
Halfter
Halfter NOM     SIN     MAS
Halfters        GEN     SIN     MAS
Halfter DAT     SIN     MAS
Halfter AKK     SIN     MAS
Halfter NOM     PLU     MAS
Halfter GEN     PLU     MAS
Halftern        DAT     PLU     MAS
Halfter AKK     PLU     MAS
@
Halfter
Halfter NOM     SIN     FEM
Halfter GEN     SIN     FEM
Halfter DAT     SIN     FEM
Halfter AKK     SIN     FEM
Halftern        NOM     PLU     FEM
Halftern        GEN     PLU     FEM
Halftern        DAT     PLU     FEM
Halftern        AKK     PLU     FEM
@
Halfter
Halfter NOM     SIN     NEU
Halfters        GEN     SIN     NEU
Halfter DAT     SIN     NEU
Halfter AKK     SIN     NEU
Halfter NOM     PLU     NEU
Halfter GEN     PLU     NEU
Halftern        DAT     PLU     NEU
Halfter AKK     PLU     NEU
```

Figure 18. Entry for *Halfter* in the Expanded Stem Lexicon.

Problems

1. Some of the words have only singular or only plural.

Figure 19 shows a list of some of the words from the Morphy lexicon having only singular or only plural. We selected only the nouns and proper nouns and grouped them by basic form in the way we structure the words in the Expanded Stem Lexicon. This is done for human readability purpose.

```
@
Gehren
Gehren  NOM     SIN     NEU
Gehrens GEN     SIN     NEU
Gehren  DAT     SIN     NEU
Gehren  AKK     SIN     NEU
```

```
@
Abreise
Abreise NOM     SIN     FEM
Abreise GEN     SIN     FEM
Abreise DAT     SIN     FEM
Abreise AKK     SIN     FEM
@
Manitu
Manitu  NOM     SIN     MAS
Manitus GEN     SIN     MAS
Manitu  DAT     SIN     MAS
Manitu  AKK     SIN     MAS
@
Kickers
Kickers NOM     PLU     NOG
Kickers GEN     PLU     NOG
Kickers DAT     PLU     NOG
Kickers AKK     PLU     NOG
@
Ehrgeiz
Ehrgeiz      NOM     SIN     MAS
Ehrgeizes    GEN     SIN     MAS
Ehrgeize     DAT     SIN     MAS
Ehrgeiz      DAT     SIN     MAS
Ehrgeiz      AKK     SIN     MAS
@
Dank
Dank NOM     SIN     MAS
Dankes       GEN     SIN     MAS
Danks GEN    SIN     MAS
Dank DAT     SIN     MAS
Danke DAT    SIN     MAS
Dank AKK     SIN     MAS
```

Figure 19. Words with no gender in the Morphy Lexicon. (grouped by basic form)

These words have only 4, 5 (see *der Ehrgeiz* above) or 6 (see *der Dank* above) forms instead of 8 or more forms. No morphological class can be induced for them in the general case and we did not tried to do so.

Special classes have to be derived for these words. For the moment we prefer to include them in neither the Ste m Lexicon nor the Expanded Stem Lexicon since this will lead to several problems. There are two solutions to this problem:

❑ *create additional morphological classes*

The new morhological classes will have rules for either only plural or only singular. In case of only plural the class will have no gender associated.

❑ *use the existing 39 classes*

Another option is to use the existing classes and assign the word *all* the classes that could cover it.

In case an unknown word of this type happens to be analysed by the present version of the System it will obtain associated all the compatible classes.

2. Some of the words do not have gender and are marked as NOG in the Morphy lexicon.

These are primarily words that have only plural forms like *die Leute*. Figure 20 shows two examples: *die Leute* and *die Bahamas*.

```
@
Leute
Leute   NOM     PLU     NOG
Leute   GEN     PLU     NOG
Leuten  DAT     PLU     NOG
Leute   AKK     PLU     NOG
@
Bahamas
```

```
Bahamas NOM     PLU     NOG
Bahamas GEN     PLU     NOG
Bahamas DAT     PLU     NOG
Bahamas AKK     PLU     NOG
```

Figure 20. Plural-only words without gender in the Morphy lexicon.

3. Some of the words are invariable and do not change.

According to the DB-MAT morphological rules it is impossible that a word does not change. Anyway, it happens that we discover some invariable forms in the Morphy lexicon. All the words of that type we investigated were due to incorrect data in the Morphy lexicon since these words actually *must* change.

```
@
Kaffee
Kaffee  NOM     SIN     MAS
Kaffee  GEN     SIN     MAS
Kaffee  DAT     SIN     MAS
Kaffee  AKK     SIN     MAS
Kaffee  NOM     PLU     MAS
Kaffee  GEN     PLU     MAS
Kaffee  DAT     PLU     MAS
Kaffee  AKK     PLU     MAS
```

Figure 21. Morphy lexicon: incorrect forms example.

Even when a word has more than 8 different forms available we sometimes face problems that prevent us from classifying it. We list here some of the main reasons for these problems:

a) There are 8 or more different forms but this is just because we have several variants for some (case,number) couple and in fact miss any information for another one. (see Figure 22)

```
@
Boxen
Boxen   NOM     SIN     NEU
Boxen   NOM     SIN     NEU
Boxens  GEN     SIN     NEU
Boxens  GEN     SIN     NEU
Boxen   DAT     SIN     NEU
Boxen   DAT     SIN     NEU
Boxen   AKK     SIN     NEU
Boxen   AKK     SIN     NEU
```

Figure 22. Morphy lexicon: incomplete forms example.

b) There are enough forms and they cover all the cases but no morphological class is able to cover them all at the same time (most likely because of incorrect lexicon data). (see Figure 23)

```
@
Bär
Bär     NOM     SIN     MAS
Bären   GEN     SIN     MAS
Bäres   GEN     SIN     MAS
Bärs    GEN     SIN     MAS
Bären   DAT     SIN     MAS
Bäre    DAT     SIN     MAS
Bär     DAT     SIN     MAS
Bär     AKK     SIN     MAS
Bären   AKK     SIN     MAS
Bären   NOM     PLU     MAS
Bären   GEN     PLU     MAS
```

```
Bären   DAT     PLU     MAS
Bären   AKK     PLU     MAS
```

Figure 23. Morphy lexicon: too many forms per word example.

c) The word has more than one gender and we failed to classify the forms for some of the genders while succeeded for another one. (see Figure 24)

```
@
Schild
Schild  NOM     SIN     NEU
Schildes        GEN     SIN     NEU
Schilds GEN     SIN     NEU
Schilde DAT     SIN     NEU
Schild  DAT     SIN     NEU
Schild  AKK     SIN     NEU
Schildern       DAT     PLU     NEU
@
Schild
Schild  NOM     SIN     MAS
Schildes        GEN     SIN     MAS
Schilds GEN     SIN     MAS
Schild  DAT     SIN     MAS
Schilde DAT     SIN     MAS
Schild  AKK     SIN     MAS
Schilde NOM     PLU     MAS
Schilde GEN     PLU     MAS
Schilden        DAT     PLU     MAS
Schilde AKK     PLU     MAS
```

Figure 24. Morphy lexicon: several genders per word example.

Figure 24 shows an example for a several genders word. We classified the second (masculine) set as *m1* but failed to do so for the first (neuter) set since there are not enough forms left.

d) The word has more than one inflection class from the same gender. We drop it. (see Figure 25)

```
@
Diakon
Diakon  NOM     SIN     MAS
Diakonen        GEN     SIN     MAS
Diakones        GEN     SIN     MAS
Diakons GEN     SIN     MAS
Diakonen        DAT     SIN     MAS
Diakone DAT     SIN     MAS
Diakon  DAT     SIN     MAS
Diakon  AKK     SIN     MAS
Diakonen        AKK     SIN     MAS
Diakone NOM     PLU     MAS
Diakonen        NOM     PLU     MAS
Diakone GEN     PLU     MAS
Diakonen        GEN     PLU     MAS
Diakonen        DAT     PLU     MAS
Diakone AKK     PLU     MAS
Diakonen        AKK     PLU     MAS
```

Figure 25. Morphy lexicon: more than one morphological class for the same gender.

Table 7 lists some statistics about the automatic morphological classes induction (Stem Lexicon and Expanded Stem Lexicon creation).

Category	**Count**
No gender	60
Multiple gender	132
Less than 8 forms	11747
Potentially good forms	13226
Classified	**13091**
Total nouns	**24975**

Table 7. Automatic stem classification statistics.

8 System Description

As has been mentioned above the System works on raw texts and its ain is the recognition and morphological classification of unknown words. This is a several stage process including:

1. Word types with unknown stem identification.
2. All possible stems generation.
3. Stem coverage refinements.
4. Morphological stems analysis.
5. Word types clusterisation.
6. Deterministic context exploitation.
7. Word types context vector creation.

We will consider these steps in more details below.

8.1 Unknown Word Tokens and Types Identification

8.1.1 What is a *word*?

Before speaking about *unknown* words we would like to define first define what is considered to be a *word* according to the System. We consider a word is a sequence of letters from the German alphabet including umlauts, the letter "ß" and the symbol "–". No number can be part of a word and no word can start with "–". A valuable discussion on the word boundaries identification can be found in (Manning & Shuetze, 1999).

8.1.2 When does a sentence end?

For the moment we consider a sentence ends if one of the following symbols occurs: ".", "!" and "?". A more sophisticated heuristic may be used later since we have a list of some important abbreviations in the Word Lexicon. (Manning & Shuetze, 1999**;** Mikheev 1999, 2000).

8.1.3 What is an *unknown word/noun/stem*?

These are three central notions: *unknown word*, *unknown noun (proper noun)* and *noun with unknown stem*. It is a question of different things and they need to be defined more precisely. An *unknown word* is a word that is missing from the Word Lexicon, while an *unknown noun* is first a *noun* and then it is noun, which is missing from the Word Lexicon. A *noun with unknown stem* is a noun whose stem is not included in the Stem Lexicon.

8.1.4 Unknown word/noun/stem identification

The unknown word tokens and types identification is a complex issue including three overlapping problems:

- ❑ *unknown words identification*
- ❑ *nouns identification*
- ❑ *nouns with unknown stem identification*

The identification of unknown word or unknown noun (we suppose we are sure it is a *noun*) is an easy problem and is solved with a single checking against the Word Lexicon. For the unknown noun identification an additional tag checking is needed to see whether the entry has either (or both) the tag SUB (noun) or the tag EIG (proper noun).

The identification of a noun with unknown stem is not so straightforward although, because in the general case we cannot be sure which is its stem in order to be able to check it against the Stem Lexicon. But we are still able to check it against the Expanded Stem Lexicon, which contains all inflected nouns (including proper nouns) that can be obtained from the known stems. Thus, there is no need to know the stem in advance. But if we find the noun in the Expanded Stem Lexicon it is sure that it has a known stem, which could be obtained by looking at the corresponding entry in the Stem Lexicon. Please, observe that a noun with unknown stem can be either known or unknown noun. It could be a known noun since the Word Lexicon can contain some nouns or pr oper nouns that are not present at the Expanded Stem Lexicon (but the reverse *cannot* happen: the Word Lexicon *must* contain all the words from the Expanded Stem Lexicon).

As has been mentioned above the System is interested in the identification and morphological classification of the nouns with unknown stems. The first thing to do is to process the text and to derive a list of all its word types. The capitalisation is discarded when deriving the list but is taken into account since for each word we derive the following three statistics:

- total frequency (TF)
- capitalised frequency (CF)
- start-of-sentence frequency (SSF)

Figure 26 shows the unknown word/noun/stem identification decision tree. The brown leaves represent the interesting cases when the word token has an unknown stem and that have to be further investigated. The first thing we try is to check the word type against the Word Lexicon, which gives us information whether the word is known or is not. In case the word is known we have to check whether it could be a noun or a proper noun, which is easily determined from the corresponding tags list in the Word Lexicon: we are looking for SUB and EIG. In case either or both of the tags SUB and EIG are present we are sure that this is a noun (a *sure noun*). The next thing to do in this case is to check the word type against the Expanded Stem Lexicon (ESL). If it is there then it is known and thus non-interesting (we know its morphological class or classes and thus there is nothing to gue ss). If it is not there then we are sure it is a noun with an unknown stem. If we suppose its stem is known and thus it is in the Stem Lexicon then the Expanded Stem Lexicon must contain all the words this stem could generate. But then the word type in question would be in the Expanded Stem Lexicon. But we checked this already and it was not there, so we get a contradiction. If the word is known but its Word Lexicon entry contains neither the SUB nor EIG tags then it cannot be a noun (a *sure non-noun* becaus e we suppose the Word Lexicon is complete in the sense that if it lists a word it contains *all* its possible POS tags) and thus is non-interesting for us.

If the word has not been found in the Word Lexicon then we can conclude it is an unknown word. In case it is a noun its stem will be unknown as well and thus will be interesting for us. On the other hand if it is a non-noun we have to skip it since we currently consider nouns only. The problem is: How to determine whether an unknown word can be a noun or not?

8.1.5 Is an unknown word a noun?

We exploit the German noun’s property to be always capitalised regardless of its position in the sentence. After the statistics above are collected we apply a simple heuristic in order to determine which of the words may be and which may not be nouns.

Heuristic

A word *cannot* be a noun iff:

1) $CF = 0$

or

2) $(SSF / CF > t)$ & $(CF < TF)$

where:

CF = (word type) Capitalised Frequency
SSF = (word type) Start of Sentence Frequency
TF = (word type) Total Frequency
t is an appropriate constant between 0 and 1. (we use 0.5)

The first condition (*CF* = 0) in the heuristic above is quite easy to understand: According to the German grammar the nouns must always be written capitalised. Thus, if there are no capitalised word tokens at all we can conclude that the word type in question cannot be a noun at least in what about the text being analysed.

The second condition is complex and not so straightforward. Let us look at the first part of the condition first (*SSF* / *CF* > *t*). It may look a bit strange: why consider the *SSF*/*CF* ratio? It is much simpler just to accept a word as a noun, when it is met capitalised in the middle of a sentence at least once. Unfortunately, this heuristic fails for a large number of cases. Table 9 shows that words like *die* appear capitalised 1000 times while it is definitely not a noun. The same applies for the subsequent highest frequency words. The *SSF*/*CF* ratio permits us to classify all these words correctly. The basic insight is that a noun is much more likely to appear in the middle of a sentence than in the beginning. Thus, if the reverse happens we can attribute the in-phrase capitalised word tokens to be due to special cases like collocations and not a manifestation of the fact that the word type could be a noun. The second part of the condition accounts for the case when all the word tokens for the word type in question are capitalised but have been always met at the beginning of a sentence. These word types are usually nouns and we included the additional condition (CF < TF) in order to do not reject them. Examples from Table 8 include: *Fleischtheke*, *Flensburg*, *Fleurs*, *Flexibilisierung*, etc.

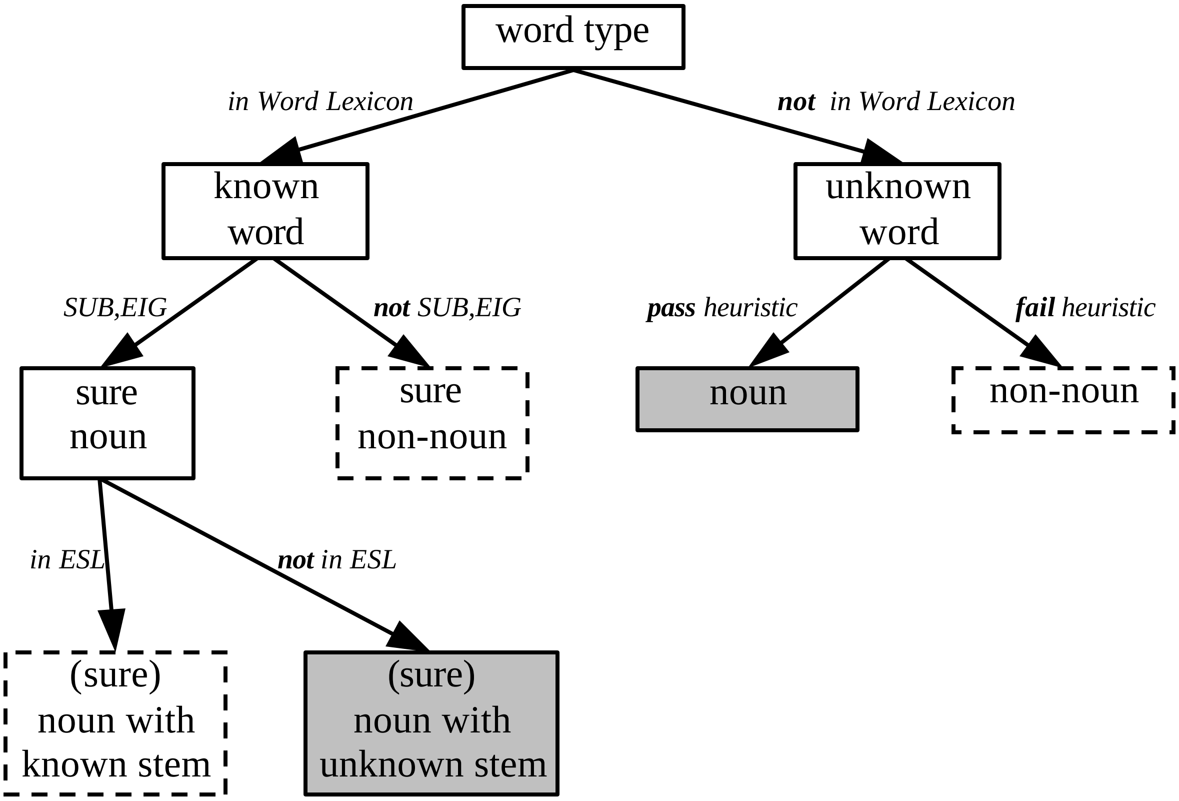


Figure 26. Unknown word/noun/stem identification decision tree

Remark

Note that some words can be used as nouns as well as other parts of speech, e.g. *erste* (Eng. "the first") can be an adjective as well as a noun (both masculine and feminine). Our heuristic does not reject them as noun candidates. We are not interested in what is *more likely* but want to process *all* the words that can be used as nouns. *Erste* for example is met 57 times in the NEGRA corpus including 10 times capitalised, 3 of which in the beginning of a sentence (which means we cannot be sure whether they would be really capitalised if used in another position in the sentence). This means that *erste* is much more likely to do *not* be a noun (with maximum likelihood estimate at least (57–10)/57=47/57). Anyway, since only 3 out of 10 capitalisations may be due to beginning of sentence we cannot reject the possibility that it may be used as noun. Thus, we keep it as noun candidate according the heuristic above.

Note as well that the heuristic sometimes fails. It considers for example the adjective *neue* (Eng. *new*) as a possible noun. We investigated the NEGRA corpus and discovered that this is due to phrases like *Verlag Neue Kritik Frankfurt a.M.*, *Neue Maxhütte* (several occurrences), *Forum Neue Musik fest* etc., where it was used as part of a capitalised collocation. A further refinement of the heuristic may include a module for automatic collocations discovery. This will help us in identifying that cases like *Neue Maxhütte* are collocations and treating their members in a different manner. We could treat these cases as start-of-sentence occurrences for example. ?

We apply the heuristic above to obtain two different lists:

1) Noun candidates list (they are likely to appear as nouns but other POS may still be possible, see Table 8)

2) List of words that *cannot* be used as nouns (at least in the text considered, see Table 9)

Accepted noun candidate	TF	CF	SSF
Jahrhunderts	7	7	6
Hans	41	41	6
Friedrich	11	11	6
Dm	17	17	6
Ziel	38	38	5
Schimmi	7	7	5
Schade	6	6	5
Peter	38	38	5
November	13	13	5
Millionen	122	122	5
Manfred	22	22	5
Kleine	43	11	5
Dieter	19	19	5
Neue	124	19	4
Munch	21	21	4
Minute	8	8	4
Michael	39	39	4
Mark	325	325	4
Männer	32	32	4
Kunst	76	76	4
Klaus	22	22	4
Kinder	105	105	4
Ina	4	4	4
Februar	7	7	4
Charles	10	10	4
Brock	21	21	4
Beispiel	40	40	4
Untersuchungen	16	16	3
Tasso	3	3	3
Sylvia	6	6	3
Ruth	6	6	3

Robert	17	17	3
Rita	5	5	3
Rainer	12	12	3
Progres	5	5	3
Petra	9	9	3
Monika	16	16	3
Michail	4	4	3
Menschen	92	92	3
Martin	13	13	3
Mal	81	25	3
Lechla	6	6	3
Lebensjahr	3	3	3
Kickers	8	8	3
Karl	15	15	3
Jürgen	15	15	3
Informationen	16	16	3
Ghozali	4	4	3
Geld	54	54	3
Geburtstag	4	4	3
Fischer	8	8	3
Fc	22	22	3
Eschbacher	7	7	3
Erste	57	10	3
Bernhard	12	12	3
Bernd	17	17	3
Bernbach	12	12	3
Barbara	5	5	3
Bad	69	67	3
April	18	18	3
Anmeldungen	5	5	3
Anmeldung	6	6	3
Anfang	43	43	3
Adorno	32	32	3
Xanana	3	3	2
Wut	3	3	2
Wolfgang	16	16	2
Waren	122	7	2
Wanzen	3	3	2
Vorsitzender	22	22	2
Vorschläge	12	12	2
Voraussetzung	6	6	2
Vergangenes	3	3	2
Vereinbarungen	4	4	2
Vectra	2	2	2
Uwe	12	12	2
Umberto	2	2	2
Ulrich	9	9	2
Ttv	4	4	2
Tsv	11	11	2
Tsg	16	16	2
Flanieren	1	1	0
Flasche	1	1	0
Flaschen	3	3	0
Flaschenpfand	1	1	0
Fleck	1	1	0
Fledermaus	6	6	0
Fledermausintrige	1	1	0
Fledermäuse	1	1	0
Fleige	2	2	0
Fleisch	2	2	0
Fleischtheke	1	1	0

Flensburg	1	1	0
Fleurs	1	1	0
Flexibilisierung	4	4	0
Flexibilität	4	4	1
Flickenteppich	1	1	0
Flickschusterei	1	1	0
Fliegenpilz	1	1	0
Flieger	1	1	0
Fliesen	1	1	0
Fließband	1	1	0
Fließwasserverbindung	1	1	0
Flirt	1	1	0
Fln	1	1	0
Fln-regimes	1	1	0
Flocki	1	1	0
Floh	1	1	0
Flohmarkt	7	7	1
Flohmarktes	1	1	0
Flohmarktstände	1	1	0
Flop	2	2	0
Flora	1	1	0
Flores	2	2	0
Florett	1	1	0
Florica	1	1	0
Florstadt	1	1	0
Floskeln	1	1	0
Flotte	1	1	0
Flower-power	1	1	0
Fluch	2	2	0
Flucht	3	3	0
Fluchtgelder	1	1	0
Flug	1	1	0
Flugblatt	1	1	0
Fluggerät	1	1	0
Fluggesellschaft	2	2	0
Fluggäste	1	1	0
Flughafen	8	8	0
Flughafenausbau	1	1	0
Flughafendienst	1	1	0
Flughafenfeuerwehr	1	1	0
Flughafenhallen	1	1	0
Flughafens	2	2	0

Table 8. NEGRA corpus: *Accepted* noun candidates list according to the heuristic, ordered by SSF.

Rejected noun candidate	TF	CF	SSF
die	5782	1000	818
der	5468	474	399
das	1721	410	350
in	2676	248	210
und	3609	149	167
im	1254	158	138
es	885	157	134
auch	925	138	133
ein	1040	151	123
doch	257	123	123
für	1242	120	109
sie	721	143	106
mit	1383	123	105

so	447	104	98
aber	388	95	89
nach	577	102	88
er	671	94	85
eine	855	100	84
wir	279	106	80
auf	1201	91	78
da	194	82	72
bei	488	77	72
am	495	85	69
wenn	279	83	66
diese	194	68	66
als	738	74	63
denn	142	59	57
von	1426	86	56
wie	533	69	54
ich	201	79	49
bis	332	24	45
was	177	53	40
wer	71	52	39
den	1898	46	39
vor	446	41	38
um	554	41	38
dabei	97	38	38
nicht	1059	40	37
seit	171	38	36
daß	531	38	35
damit	127	35	34
nur	470	33	31
man	225	41	31
dann	190	32	31
aus	645	41	31
noch	514	35	30
zu	1423	32	29
zum	401	36	28
dies	61	27	27
während	72	31	26
unter	205	31	25
an	689	30	25
jetzt	130	26	24
alle	184	28	24
schon	252	29	23
dazu	89	24	22
oder	329	20	21
respektierend	1	0	0
respektvoll	1	0	0
restlichen	2	0	0
resultieren	1	0	0
resultierende	1	0	0
resultiert	1	0	0
resümiert	2	0	0
resümierte	1	0	0
rethorisch	1	0	0
retten	5	0	0
rettende	1	0	0
rettender	1	0	0
rettete	3	0	0
revidierte	1	0	0
revolutionäre	2	0	0
revolutionären	1	0	0
rezitierte	1	0	0

rheinländischen	1	0	0
rhetorisch	1	0	0
rhythmisch	2	0	0
rhythmische	1	0	0
rhythmusorientierter	1	0	0
ric	2	1	1
richten	1	0	0
richtet	4	0	0
richtete	5	0	0
richteten	1	0	0
richtig	20	3	2
richtige	9	0	0
richtigen	7	0	0
rieben	1	0	0
riecht	1	0	0
rief	4	0	1
riesenhaften	1	0	0
riesige	2	0	0
riesigen	2	0	0
riesiger	1	0	0
riesiges	2	0	0
rigidem	1	0	0
rigoros	1	0	0
rigorosen	1	0	0
ringen	1	0	0
ringt	3	0	0
riskant	1	0	0
riskante	1	0	0
riskanten	2	0	0
riskantes	1	0	0
riß	1	0	0
rohstoffarmes	1	0	0
rollstuhlgerechte	1	0	0
rollten	3	0	0
romanischen	2	0	0
romantische	2	0	0

Table 9. NEGRA corpus: *Rejected* noun candidates list according to the heuristic, ordered by SSF.

The heuristic above discovered 19066 candidates for nouns and rejected 10379.

Remark

A much simpler heuristic is possible: *If a word is met capitalised in the middle of a sentence it is considered to be a potential noun.* Note that it will fail to identify correctly as non-nouns words like *die, der* etc. (see Table 9). In plus the automatic discovery and taking into account the collocations seems much more necessary. Perhaps one would like to add them to the words that appear only in the beginning of a sentence. Both heuristics have to be tested against either/both subsets of the NEGRA corpus and a collection of raw texts of different sizes using the Morphy lexicon. ?

The System outputs the results of this step in three files as follows:

- *non-nouns*

  The non-nouns are output in the *non-nouns file* (see Table 10). The file is sorted aplhabetically and each line contains a single word type followed by *TF*, *CF* and *SSF*. The word *LEXICON* which may appear in the last column shows that the word has been found in the Word Lexicon but neither of the POS tags *SUB* nor *EIG* was present there and thus the conclusion that it cannot be a noun has been derived. The statistics *TF*, *CF* and *SSF* can help understand why a word type has been decided to be a non-noun (neither common nor proper). Remember that although these statistics have been output for all the word types listed they were really taken into account *only* if the word type has not been found in the Word Lexicon.

- *nouns*
  - *nouns with known stem*

    The nouns with known stem are output in the *nouns with* ***known*** *stem file* (see Table 11). The file is sorted alphabetically and each line contains a single noun followed by a list of its possible stems. If more than one stem is possible they are all listed there separated by a comma. Each stem is followed by a list of its morphological classes enclosed in parentheses. All the word types listed there are nouns and have been found in the Expanded Stem Lexicon. Remember that according to the lexicons' construction this means that their stems are known and could be found in the Stem Lexicon.
  - *nouns with unknown stem*

    The nouns with unknown stem are output in the *nouns with* ***unknown*** *stem file* (see Table 12). The file is sorted aplhabetically and each line contains a single word type followed by *TF*, *CF* and *SSF*. The word *LEXICON* which may appear in the last column shows that the word has been found in the Word Lexicon with at least one of the POS tags *SUB* or *EIG*, but has not been found in the Expanded Stem Lexicon. This means that its stem was unknown although the word type itself is known and is included in the Word Lexicon. The statistics *TF*, *CF* and *SSF* can help understand why a word type has been decided to be a noun (either common or proper noun). Remember that although these statistics have been output for all the word types listed they were really taken into account *only* if the word type has not been found in the Word Lexicon.

Non-noun	TF	CF	SSF	Found in the Word Lexicon?
....	....	...	...	....
rekognoszieren	6	0	0	
rekognoszierenden	1	0	0	
rekognosziert	1	0	0	
rekrutierte	1	0	0	LEXICON
relativ	1	0	0	LEXICON
ren	1	0	0	
rennt	3	0	0	LEXICON
repetieren	1	0	0	
repräsentieren	1	0	0	LEXICON
republikanisch	2	0	0	
republikanische	1	0	0	
republikanischen	1	0	0	
requirierten	1	0	0	LEXICON
reserviert	1	0	0	LEXICON
respektablen	1	0	0	LEXICON
respektvoll	1	0	0	LEXICON
respektvollen	1	0	0	LEXICON
rette	2	1	1	LEXICON
retten	17	0	0	LEXICON
rettet	2	1	1	LEXICON
rettete	2	0	0	LEXICON
rettungslos	1	0	0	
richten	8	0	0	LEXICON
richterlichen	1	0	0	LEXICON
richtest	1	0	0	LEXICON
richtet	5	1	1	LEXICON
richtete	17	0	0	LEXICON
richteten	7	0	0	LEXICON
richtig	23	2	2	LEXICON
richtigen	12	0	0	LEXICON
richtiger	3	0	0	LEXICON
richtigere	1	0	0	LEXICON

richtiges	1	0	0	LEXICON
rieb	3	0	0	LEXICON
riechen	3	0	0	LEXICON
riecht	1	0	0	LEXICON
rief	121	0	66	LEXICON
riefen	4	0	3	LEXICON
riegelt	1	1	1	LEXICON
riesenhaft	1	0	0	LEXICON
riesenhafte	1	0	0	LEXICON
riesenkräftigen	1	0	0	
riesenstarken	1	0	0	
riesige	2	0	0	LEXICON
riesigen	6	0	0	LEXICON
riet	3	0	0	LEXICON
rietest	1	0	0	LEXICON
ringenden	1	0	0	LEXICON
ringender	1	0	0	LEXICON
ringsum	5	0	0	
rinnenden	1	0	0	LEXICON
rinnt	2	0	0	LEXICON
riskieren	2	0	0	LEXICON
riskiert	1	0	0	LEXICON
roch	2	0	0	LEXICON
rohen	2	0	0	LEXICON
rollte	1	0	0	LEXICON
romantischer	1	0	0	LEXICON
rostbedeckten	1	0	0	
roter	24	6	0	LEXICON
rotes	3	0	0	LEXICON
rothaarigen	1	0	0	LEXICON
rothäutigen	2	0	0	
rotwangiges	1	0	0	
rotwollenes	1	0	0	
ruchlose	1	0	0	
....	....	...	...	....

Table 10. *Non-nouns file* (extraction). The file is sorted aplhabetically and each line contains a single word type followed by TF, CF and SSF. The word "LEXICON" which may appear in the last column shows that the word has been found in the Word Lexicon but neither of the POS tags SUB nor EIG was found there.

Noun with Known Stem	Stems and morphological classes
Deutsche	Deutsche(m7 f19)
Deutschen	Deutsche(m7 f19)
....	....
Eskorte	Eskorte(f16)
Etikette	Etikett(n20), Etikette(f16)
....	....
Leiter	Leiter(m4 f16)
Leitern	Leiter(m4 f16)
....	....
Halfter	Halfter(m4 f16 n23)
....	....
Herberge	Herberge(f16)
Herbst	Herbst(m1)
Herde	Herd(m1), Herde(f16)
Herden	Herd(m1), Herde(f16)
....	....
Recht	Recht(n20)
Rechte	Recht(n20), Rechte(f16)
Rechten	Recht(n20), Rechte(f16)
Rechts	Recht(n20)
Recken	Recke(m7)

```
.............   .................................
Requisiten      Requisit( n25 ), Requisite( f16 )
Rest            Rest( m1 )
Restaurants     Restaurant( n24 )
Reste           Rest( m1 )
Resultat        Resultat( n20 )
Resultate       Resultat( n20 )
Retter          Retter( m4 )
Rettung         Rettung( f17 )
Revision        Revision( f17 )
Revolver        Revolver( m4 )
Revolvern       Revolver( m4 )
Revolvers       Revolver( m4 )
Richter         Richter( m4 )
Richtung        Richtung( f17 )
Richtungen      Richtung( f17 )
Riechorgan      Riechorgan( n20 )
Riechorgane     Riechorgan( n20 )
Riegel          Riegel( m4 )
Riemen          Riemen( m4 )
Riese           Riese( m7 )
Riesen          Riese( m7 )
Rind            Rind( n21 )
Rinde           Rind( n21 ), Rinde( f16 )
```

Table 11. *Nouns with known stem file* (extractions). The file is sorted alphabetically and each line contains a single noun followed by a list of its possible stems. If more than one stem is possible they are all listed there. Each stem is followed by a list of its morphological classes in parentheses. All the word types listed there are nouns and have been found in the Expanded Stem Lexicon.

Noun with Unknown Stem	TF	CF	SSF	Found in the Word Lexicon?
....	....	....	....	....
Danke	16	12	11	LEXICON
Dankesadresse	1	1	0	
Dasein	1	1	0	LEXICON
Dauben	1	1	0	
Dauerlaufe	1	1	0	
David	1	1	0	LEXICON
Davis	8	8	2	
Davonkommen	2	1	0	LEXICON
Davonreitenden	1	1	0	
Death	379	379	0	
Deaths	23	23	0	
Deck	5	5	0	
Deckenlücke	1	1	0	
Deckes	1	1	0	
Deckhands	3	3	0	
Dehors	1	1	0	
Detachement	4	4	0	
Detachements	3	3	0	
Deutlichkeit	2	2	0	LEXICON
Deutschland	1	1	0	LEXICON
Deutschlands	1	1	0	LEXICON
Deutschtums	1	1	0	LEXICON
Diagonale	1	1	0	LEXICON
Dichtheit	1	1	0	
Dichtkunst	1	1	0	
Dickhornschaf	1	1	0	
Dickicht	4	4	0	
Dickschwanz	3	3	0	
Dickschwanzfelle	1	1	0	

Diebstahlsgeschichte	1	1	0	
Dienerschaft	1	1	0	
Dietrich	3	3	0	LEXICON
Dietrichs	1	1	0	LEXICON
Digger	1	1	0	
Diggins	3	3	0	
Directory	1	1	0	
Diskretion	1	1	0	LEXICON
Distinktion	1	1	0	
Donnerschlag	1	1	0	
Donnerworte	1	1	0	
Doppelbüchse	2	2	0	
Doppelbüchsen	1	1	0	
Doppelgewehr	2	2	0	
Doppelmord	1	1	0	
Doppelpistol	1	1	0	
Doppelplan	1	1	0	
Doppelpässe	1	1	0	
Dragonersergeant	1	1	0	
Drange	1	1	0	
Draußenstehenden	1	1	0	
Drehbank	1	1	0	
Drehpistole	1	1	0	
Dreien	9	5	0	
Dreihundert	2	2	2	
Dreispitzhut	1	1	0	
Dreißiger	1	1	0	
....	....	....	....	....

Table 12. *Nouns with unknown stem file* (extraction). The file is sorted aplhabetically and each line contains a single word type followed by TF, CF and SSF. The word "LEXICON" which may appear in the last column shows that the word has been found in the Word Lexicon with at least one of the POS tags SUB or EIG, but has not been found in the Expanded Stem Lexicon.

8.2 All possible stems generation

8.2.1 Rules and generation

We go through the words and generate all the possible stems that can be obtained by reversing all acceptable German inflexions for the word type while taking in account the umlauts and the *ß* alternations. Table 13 lists all possible inflexion rules we are trying to reverse. The inflexion rules are derived from the rules in the 39 morphological classes that we use. In fact some of these are equivalent: `[se]` for example is either `0` or `se`.

0	"	"e	"en	"er	"ern	"n	"n(2)	[e]	[e]s	[e]s(1)
[r]	[s]	[se]	a	as	e	en	er	ern	es	ien
n	n(2)	nen	ns	s	se	sen	ses	um	ums	us

Table 13. All distinct inflexion rules applied by the morphological classes.

For each word type all acceptable rule inversions are performed. For example for the word *Lehrerinnen* the following stems are generated (by removing *-nen*, *-en*, *-n* and *0*):

Lehrerin, *Lehrerinn*, *Lehrerinne*, *Lehrerinnen*

We do not impose any limitations when generating a stem except that it must be at least one character long.

Remark

One may argue that imposing at least *l*=3 characters stem length and at least one vowel is a reasonable limitation. Unfortunately, this will result in missing some common two-character abbreviations e.g. *DM*, *AG*, *CD* or *TU*. Even the letters of the alphabet can be useful stems (although just one-letter long) e.g. *A* or *O*. These letters are really used as words in German and are included in the Morphy lexicon and from there — in our Word Lexicon.

Of course all the letters of the alphabet could be included in the lexicon but this would hardly make any sense for the two-character letters. Anyway, allowing at least two-letter stems may be reasonable limitation but for the moment we prefer to keep the stem generation process limitation free. ?

8.2.2 Comments and examples

The purpose of the stem generation process is to both identify all the acceptable stems and group the inflected forms of the same word together. For this purpose we remember all the word types that generated the stem. If we manage to perform the right stemming then all its corresponding inflected word type forms present in the analysed text will be grouped together (see Table 14, Table 15, Table 16). We would like to stress that although the different word types that are inflected forms of the same word will be grouped under the same stem there may be some additional word types. They belong to another stem but under certain rules they are able to generate the current one as well. Let us take for example the first row from Table 14. The stem *Haus* is the correct stem for the word *das Haus*, whose morphological class is *n22*. All the word forms listed there are correct except *Hausse* and *Hausen*. The latter are valid candidates for this stem according to the rules from Table 13 but are incompatible with the correct morphological class *n22*. We will not try to resolve these problems at this stage and will return to them later. What is important for now is that:

- *We have all the possible stems that could be obtained by reversing the rules.*
- *The inflected forms of the same word are grouped together given the correct stem.*

Remark

Table 14, Table 15 and Table 16 show the head, the middle and the tail of the stems list for *all* the word types from the NEGRA corpus ordered by word types covered by the stem count. We did so just for illustration purposes. Normally the System applies the stem generation to *word types with unknown stem only*. ?

Stem	#	Word types covered
Haus	**7**	**{ Haus, Hause, Hausen, Hauses, Hausse, Häuser, Häusern }**
Groß	6	{ Große, Großen, Großer, Großes, Größe, Größen }
Große	6	{ Große, Großen, Großer, Großes, Größe, Größen }
Spiel	**6**	**{ Spiel, Spiele, Spielen, Spieler, Spielern, Spiels }**
Ton	**6**	**{ Ton, Tonnen, Tons, Tonus, Töne, Tönen }**
Band	5	{ Band, Bandes, Bände, Bänder, Bändern }
Bau	5	{ Bau, Bauen, Bauer, Bauern, Baus }
Beruf	5	{ Beruf, Berufe, Berufen, Berufes, Berufs }
Besuch	5	{ Besuch, Besuchen, Besucher, Besuchern, Besuches }
Brief	5	{ Brief, Briefe, Briefen, Briefes, Briefs }
Erfolg	5	{ Erfolg, Erfolge, Erfolgen, Erfolges, Erfolgs }
Fall	5	{ Fall, Falle, Falles, Fälle, Fällen }
Geschäft	5	{ Geschäft, Geschäfte, Geschäften, Geschäftes, Geschäfts }
Grund	5	{ Grund, Grunde, Gründe, Gründen, Gründer }
Hau	5	{ Hau, Haus, Hause, Hausen, Hauses }
Jung	5	{ Jung, Junge, Jungen, Junger, Jungs }
Kampf	5	{ Kampf, Kampfes, Kämpfe, Kämpfen, Kämpfer }
Kur	5	{ Kur, Kurs, Kurse, Kursen, Kurses }
Kurs	5	{ Kurs, Kurse, Kursen, Kurses, Kursus }

Land	5	{ Land, Lande, Landes, Länder, Ländern }
Lauf	5	{ Lauf, Laufe, Laufen, Läufe, Läufer }
Mann	5	{ Mann, Mannen, Mannes, Männer, Männern }
Ortsbeirat	5	{ Ortsbeirat, Ortsbeirates, Ortsbeirats, Ortsbeiräte, Ortsbeiräten }
Roll	5	{ Roll, Rolle, Rollen, Roller, Rolls }
Sach	5	{ Sache, Sachen, Sacher, Sachs, Sachsen }
Sieg	5	{ Sieg, Siegen, Sieger, Sieges, Siegs }
Stein	5	{ Stein, Steine, Steinen, Steiner, Steines }
Stück	5	{ Stück, Stücke, Stücken, Stückes, Stücks }
Treff	5	{ Treff, Treffen, Treffer, Treffern, Treffs }
Verein	5	{ Verein, Vereine, Vereinen, Vereines, Vereins }
Volk	5	{ Volk, Volker, Volkes, Völker, Völkern }
Feld	4	{ Feld, Felder, Feldern, Feldes }
Film	4	{ Film, Filme, Filmen, Films }
Frankfurt	4	{ Frankfurt, Frankfurter, Frankfurtern, Frankfurts }
Freund	4	{ Freund, Freunde, Freunden, Freundes }
Geld	4	{ Geld, Gelder, Geldern, Geldes }
Gemeindehaushalt	4	{ Gemeindehaushalt, Gemeindehaushalte, Gemeindehaushaltes, Gemeindehaushalts }
Gesicht	4	{ Gesicht, Gesichter, Gesichtern, Gesichts }
Grau	4	{ Grau, Graue, Grauen, Graus }
Grun	4	{ Grün, Grüne, Grünen, Grüner }
Grün	4	{ Grün, Grüne, Grünen, Grüner }
Gut	4	{ Gute, Guten, Gutes, Güter }
Handel	4	{ Handel, Handeln, Handels, Händel }
Hoh	4	{ Hohen, Hohes, Höhe, Höhen }
Hohe	4	{ Hohen, Hohes, Höhe, Höhen }
Institut	4	{ Institut, Institute, Instituten, Instituts }
Instrument	4	{ Instrument, Instrumente, Instrumenten, Instruments }
International	4	{ International, Internationale, Internationalen, Internationales }
Italien	4	{ Italien, Italiener, Italienern, Italiens }
Jahr	4	{ Jahr, Jahre, Jahren, Jahres }
Jahrhundert	4	{ Jahrhundert, Jahrhunderte, Jahrhunderten, Jahrhunderts }
Jo	4	{ Joe, Jon, Jos, Jose }
Kind	4	{ Kind, Kinder, Kindern, Kindes }
Kinderarzt	4	{ Kinderarzt, Kinderarztes, Kinderärzte, Kinderärzten }
Konflikt	4	{ Konflikt, Konflikte, Konflikten, Konfliktes }
Konzert	4	{ Konzert, Konzerte, Konzerten, Konzerts }
Krei	4	{ Kreis, Kreise, Kreisen, Kreises }
Kreis	4	{ Kreis, Kreise, Kreisen, Kreises }
Krieg	4	{ Krieg, Krieger, Kriegern, Krieges }
Kunstwerk	4	{ Kunstwerk, Kunstwerke, Kunstwerken, Kunstwerks }
Lang	4	{ Lang, Lange, Langen, Länge }
Lebensjahr	4	{ Lebensjahr, Lebensjahren, Lebensjahres, Lebensjahrs }
Mal	4	{ Mal, Male, Malen, Maler }
Mitglied	4	{ Mitglied, Mitglieder, Mitgliedern, Mitglieds }
Monat	4	{ Monat, Monate, Monaten, Monats }
Mord	4	{ Mord, Morde, Morden, Mörder }
Motiv	4	{ Motiv, Motive, Motiven, Motivs }
Neu	4	{ Neue, Neuen, Neuer, Neues }
Neue	4	{ Neue, Neuen, Neuer, Neues }
Not	4	{ Not, Note, Noten, Nöte }
Ortsbezirk	4	{ Ortsbezirk, Ortsbezirke, Ortsbezirken, Ortsbezirks }
Ost	4	{ Ost, Osten, Oster, Ostern }
Parlament	4	{ Parlament, Parlamente, Parlamenten, Parlaments }
Plan	4	{ Plan, Planer, Pläne, Plänen }
Platz	4	{ Platz, Platzes, Plätze, Plätzen }
Politik	4	{ Politik, Politiker, Politikern, Politikum }
Problem	4	{ Problem, Probleme, Problemen, Problems }

Programm	4	{ Programm, Programme, Programmen, Programms }
Raum	4	{ Raum, Raumes, Räume, Räumen }
Recht	4	{ Recht, Rechte, Rechten, Rechts }
Rei	4	{ Rein, Reis, Reise, Reisen }
Schul	4	{ Schule, Schulen, Schüler, Schülern }
Schult	4	{ Schulte, Schulter, Schultern, Schultes }
Schutz	4	{ Schutz, Schütz, Schütze, Schützen }
Sohn	4	{ Sohn, Sohnes, Söhne, Söhnen }
Stadtteil	4	{ Stadtteil, Stadtteile, Stadtteilen, Stadtteils }
Stand	4	{ Stand, Stande, Stände, Ständen }
Standort	4	{ Standort, Standorte, Standorten, Standortes }
Steine	4	{ Steine, Steinen, Steiner, Steines }
Studi	4	{ Studie, Studien, Studium, Studiums }
Sturm	4	{ Sturm, Sturmes, Stürmen, Stürmer }
System	4	{ System, Systeme, Systemen, Systems }
Säugling	4	{ Säugling, Säuglinge, Säuglingen, Säuglings }
Tag	4	{ Tag, Tage, Tagen, Tages }
Tanz	4	{ Tanz, Tänze, Tänzer, Tänzern }
Termin	4	{ Termin, Termine, Terminen, Terminus }
Tisch	4	{ Tisch, Tische, Tischen, Tisches }
To	**4**	**{ To, Ton, Tons, Tor }**
Turnier	4	{ Turnier, Turniere, Turnieren, Turniers }
Umsatz	4	{ Umsatz, Umsatzes, Umsätze, Umsätzen }
Verband	4	{ Verband, Verbandes, Verbände, Verbänden }
Verhältnis	4	{ Verhältnis, Verhältnisse, Verhältnissen, Verhältnisses }
Versuch	4	{ Versuch, Versuche, Versuchen, Versuchs }
Vorstand	4	{ Vorstand, Vorstandes, Vorstands, Vorstände }
Wand	4	{ Wand, Wandern, Wände, Wänden }
Weg	4	{ Weg, Wege, Weges, Wegs }
Wei	4	{ Wein, Weinen, Weise, Weisen }
Werk	4	{ Werk, Werke, Werken, Werkes }
West	4	{ West, Weste, Westen, Western }
Wie	4	{ Wien, Wiens, Wiese, Wiesen }
Wort	4	{ Wort, Worte, Worten, Wortes }
Zahl	4	{ Zahl, Zahlen, Zähler, Zählern }
Zug	4	{ Zug, Zuge, Züge, Zügen }
Zweig	4	{ Zweig, Zweige, Zweigen, Zweigs }

Table 14. Largest coverage stems (NEGRA corpus) ordered by word types covered count.

Stem	#	Word types covered
Tonband	2	{ Tonband, Tonbändern }
Tone	**2**	**{ Töne, Tönen }**
Tonn	**2**	**{ Tonne, Tonnen }**
Tonne	**2**	**{ Tonne, Tonnen }**
Tore	2	{ Tore, Toren }
Torhau	2	{ Torhaus, Torhauses }
Torhaus	2	{ Torhaus, Torhauses }
Tour	2	{ Tour, Touren }
Tourist	2	{ Tourist, Touristen }
Tourne	2	{ Tournee, Tourneen }
Tournee	2	{ Tournee, Tourneen }
Traditio	2	{ Tradition, Traditionen }
Tradition	2	{ Tradition, Traditionen }
Trag	2	{ Tragen, Träger }
Trainer	2	{ Trainer, Trainers }
Transport	2	{ Transport, Transporte }
Traum	2	{ Traum, Träume }

Trebur	2	{ Trebur, Treburer }
Treffe	2	{ Treffen, Treffer }
Treffer	2	{ Treffer, Treffern }
Trinkwasser	2	{ Trinkwasser, Trinkwassers }
Trockn	2	{ Trocknen, Trockner }
Trockne	2	{ Trocknen, Trockner }
Tropf	2	{ Tropf, Tropfen }
Tropha	2	{ Trophäe, Trophäen }
Trophae	2	{ Trophäe, Trophäen }
Trophä	2	{ Trophäe, Trophäen }
Trophäe	2	{ Trophäe, Trophäen }
Trupp	2	{ Truppe, Truppen }
Truppe	2	{ Truppe, Truppen }
Tun	2	{ Tun, Tuns }
Turin	2	{ Turin, Turiner }
Turk	2	{ Türke, Türken }
Turke	2	{ Türke, Türken }
Turn	2	{ Turnen, Turner }
Turne	2	{ Turnen, Turner }
Turniere	2	{ Turniere, Turnieren }
Tön	**2**	**{ Töne, Tönen }**
Töne	**2**	**{ Töne, Tönen }**
Töpf	2	{ Töpfe, Töpfer }
Töpfe	2	{ Töpfe, Töpfer }
Türk	2	{ Türke, Türken }
Türke	2	{ Türke, Türken }
Türr	2	{ Türr, Türrs }
U	2	{ U, Un }
U-bah	2	{ U-bahn, U-bahnen }
U-bahn	2	{ U-bahn, U-bahnen }

Table 15. In the middle of the sorted stems list (NEGRA corpus) ordered by word types covered count.

Stem	#	Word types covered
A	1	{ A }
A-	1	{ A- }
Abad	1	{ Abad }
Abbau	1	{ Abbau }
Abberufung	1	{ Abberufung }
Abbilde	1	{ Abbilder }
Abbilder	1	{ Abbilder }
Abbildung	1	{ Abbildung }
Abbruch	1	{ Abbruch }
Abd	1	{ Abd }
Abdesalaam	1	{ Abdesalaam }
Abenden	1	{ Abenden }
Abendes	1	{ Abendessen }
Abendess	1	{ Abendessen }
Abendesse	1	{ Abendessen }
Abendessen	1	{ Abendessen }
Abendmusik	1	{ Abendmusik }
Abendroth	1	{ Abendroth }
Abendschul	1	{ Abendschule }
Abendschule	1	{ Abendschule }
Abendwind	1	{ Abendwind }
Abenteu	1	{ Abenteuer }
Abenteue	1	{ Abenteuer }
Abenteuer	1	{ Abenteuer }

Abenteuerrei	1	{ Abenteuerreise }
Abenteuerreis	1	{ Abenteuerreise }
Abenteuerreise	1	{ Abenteuerreise }
Abenteuer-verei	1	{ Abenteuer-vereinen }
Abenteuer-verein	1	{ Abenteuer-vereinen }
Abenteuer-vereine	1	{ Abenteuer-vereinen }
Abenteuer-vereinen	1	{ Abenteuer-vereinen }
Abenteur	1	{ Abenteurer }
Abenteure	1	{ Abenteurer }
Abenteurer	1	{ Abenteurer }
Aberwitz	1	{ Aberwitzes }
A-jugend	1	{ A-jugend }
A-jugend-turni	1	{ A-jugend-turnier }
A-jugend-turnie	1	{ A-jugend-turnier }
A-jugend-turnier	1	{ A-jugend-turnier }
A-klas	1	{ A-klasse }
A-klass	1	{ A-klasse }
A-klasse	1	{ A-klasse }
A-landerspiel	1	{ A-länderspiele }
A-landerspiele	1	{ A-länderspiele }
A-lauf	1	{ A-lauf }
A-länderspiel	1	{ A-länderspiele }
A-länderspiele	1	{ A-länderspiele }
A-promotio	1	{ A-promotion }
A-promotion	1	{ A-promotion }
A-waff	1	{ A-waffen }
A-waffe	1	{ A-waffen }
A-waffen	1	{ A-waffen }
A-waffen-trag	1	{ A-waffen-träger }
A-waffen-trager	1	{ A-waffen-träger }
A-waffen-träg	1	{ A-waffen-träger }
A-waffen-träge	1	{ A-waffen-träger }
A-waffen-träger	1	{ A-waffen-träger }

Table 16. The end of the sorted stems list (NEGRA corpus) ordered by word types covered count.

8.3 Stem coverage refinements

8.3.1 Analysis

Before we answer the question, which is the morphological class for a word with an unknown stem we must answer the more fundamental one: Which of the word tokens observed in the analysed raw text are forms of the same word? Despite the need to know what the unknown words are actually, having several inflected forms of an unknown word implies several constraints and gives an important information about its possible morphological class while helping the identification of the corresponding stem.

By generating all the possible stems we made the first step in the direction of both word forms grouping and stem identification: we have a list of *all* acceptable stems that could generate the word types observed and there is always at least one stem that groups together the inflected forms of the same word. But these results have to be refined further. As have been mentioned above there are a lot of false stems and even the correct ones cover some word types that actually have a different stem. We illustrated this with the word *das Haus*. Let us take another example from Table 14: *das Spiel*. This word actually has a known stem and its morphological class is *n20*. There are 6 different word forms listed that could stem to *Spiel*: *Spiel*, *Spiele*, *Spielen*, *Spieler*, *Spielern* and *Spiels*. Looking at the endings the morphological class *n20* can take (see Table 3) we find that the forms *Spielern* and *Spieler* are invalid. In fact they are inflected forms of another word: *der Spieler* with morphological class *m4* and stem *Spieler*.

Consider the stem *Ton* as another example. Looking at Table 14 we see it is supposed to cover *Ton, Tonnen, Tons, Tonus, Töne* and *Tönen*. The word *der Ton* has a known stem *Ton* and morphological class *m4*. This means that the forms *Tonnen* and *Tonus* are attached there incorrectly. Looking below in the same Table 14 we find the stem *To* covering *Ton* and *Tons* that are in fact forms of *der Ton* and two additional word types: *To* and *Tor*. Investigating Table 15 we discover three related stems more: *Tone*, *Tonn* and *Tonne*. The stem *Tone* covers the forms *Töne*, *Tönen* (in fact forms of *der Ton*), while both *Tonn* and *Tonne* cover the forms *Tonne* and *Tonnen*. The stem *Tonn* is wrong and the stem *Tonne* is correct with morphological class *f16*. *Die Tonne* covers correctly the invalid form *Tonnen* we found while looking at the stem *Ton*. Further in the file (outside the file fragments shown in the tables) appears a good stem for the form *Tonus*: the stem *Tonus* with morphological class *m11*.

Thus, it happens that we have false but still possible stems (like *Tonn*) and some of the correct stems cover some invalid words (like *Ton*). But there are still perfectly good stems that are both valid and cover only correct word forms (like *Tonne*). Because of the way we generated the stems and their coverage in the previous step we can be sure in cases like *Tonne* that ***all*** the correct forms of the stem present in the text are covered.

What we would like to have before trying to find the stem morphological class(es) is a set of valid stems each one covering only valid and all the valid words found in the text. We solve the problem in two steps. In the first step we refine the stem coverage in a way that the stem covers only "compatible" word forms in the sense that there exists at least one morphological class that could generate all these forms given the stem. In the second step we select some of the stems and reject the others in a way that each word is covered by exactly one stem. We will explain the process in more details below.

8.3.2 Refinement

We start with the stem coverage refinements. We go through the stems and for each one we check whether there exists a morphological class that could generate all the word forms. If at least one is found we accept the current coverage and otherwise we try to refine it in order to make it acceptable. As we saw above it is possible that a stem may be generated by a set of words that it cannot cover together. It is important to say that at this moment we are *not* interested in the question whether this stem is really correct but just in whether it is compatible with all the word forms it covers taken together. As an example that a stem can be incorrect consider the word form *Tages*. According to our stem generation strategy from the previous section the following stems will be generated: *Tages*, *Tage* and *Tag*. While all the three stems are valid since they have been obtained by reversing only legal rules from Table 13, there is exactly one correct stem: *Tag*.

How to refine the class? An obvious (but not very wise) solution is just to reject the stem. But we are not willing to do so since this may result in losing a useful stem. We do not have to reject the stem *Spiel* for example just because it is incompatible with the set of words it covers taken together. But anyway, suppose the stem *Spiel* is unknown. How could we then decide that *Spiel*, *Spiele*, *Spielen* and *Spiels* are correct, while *Spieler*, *Spielern* are not and must be rejected? The first group is covered by the classes *m1*, *m9* (and *m9a* that has been conflated to *m9*), *n20* and *n25*, while the second is covered by *m3a* and *n21*. Thus, both groups are acceptable. What could make us decide that *Spiel* is not the correct stem for *Spieler* and *Spielern*, while there are two morphological classes that can generate these forms? And if we have to choose between the two groups why will we reject the latter? The obvious answer is simply because the first group is bigger and thus it is more likely to be correct.

What is important here is that we *choose* between the two groups. By doing so we presuppose that the stem *Spiel* has exactly one morphological class. Otherwise we could accept both groups together with all acceptable word forms subsets that could be covered by a rule. This obviously leads to combinatorial expansion of the possibilities to be considered and makes the model much more complex than necessary. In fact it is quite unlikely that a word has more than one morphological class: the Stem Lexicon contains only 73 such stems, all listed in Figure 11, out of

13,147 stems. In our opinion, it is even more unlikely that a new unknown word first, has more than one morphological class, and in plus is used with two or more of these classes at the same text. We thus always look for only one word form set possibility for the stem coverage given the stem. And we always prefer the biggest word forms set that a morphological class could cover.

An interesting issue is the case when we have more than one candidate for the same stem. Let us take for example the stem *Schrei*, which is generated by three words: *Schrei*, *Schreien* and *Schreier*.

It can cover no more than 2 of these at the same time: either {*Schrei*, *Schreien*} or {*Schrei*, *Schreier*}. How to choose between the two options? The simplest solution again is just to reject the stem, in which case we obtain that all the 3 word types are unrelated and each one forms its own stem while the correct choice is the further one. We solve the problem by keeping the set, which is most likely.

How do we decide which set is more likely in case they have the same elements number? We select the one that is covered by the more likely morphological class. We estimated the probability for each morphological class from 8,5MB of raw German texts. We extracted the word types and checked them against the Expanded Stem Lexicon. If the word type was there we extracted its stem(s) from the Stem Lexicon and increased the frequency of the corresponding morphological class(es) by the current word type frequency. The frequencies obtained are listed in Table 17, Table 18 and Table 19 together with the corresponding maximum likelihood probability estimations (in %). Since all morphological classes frequencies are distinct this test is well-defined and always designates a single winner.

Class	**Count**	**% from masculine**	**% from total**
m1	*21,400*	28.962%	11.388%
m1a	*28*	0.038%	0.015%
m2	*14,544*	19.683%	7.740%
m3	*2,973*	4.024%	1.582%
m3a	*97*	0.131%	0.052%
m4	*14,808*	20.041%	7.880%
m5	*2,309*	3.125%	1.229%
m6	*7,046*	9.536%	3.750%
m7	*6,537*	8.847%	3.479%
m7a	*628*	0.850%	0.334%
m8	*2,665*	3.607%	1.418%
m9	*701*	0.949%	0.373%
m10	*144*	0.195%	0.077%
m11	*10*	0.014%	0.005%
Total	***73,890***	**100.00%**	**39.32%**

Table 17. Masculine: Class frequency and maximum likelihood estimation probability (raw text using the lexicons).

Class	**Count**	**% from feminine**	**% from total**
f12	*127*	0.175%	0.068%
f13	*235*	0.323%	0.125%
f14	*5,187*	7.133%	2.760%
f14a	*651*	0.895%	0.346%
f15	*465*	0.639%	0.247%
f15a	*79*	0.109%	0.042%
f16	*36,536*	50.245%	19.443%
f17	*28,432*	39.101%	15.130%

f18	*761*	1.047%	0.405%
f19	*242*	0.333%	0.129%
Total	***72,715***	**100.00%**	**38.70%**

Table 18. Feminine: Class frequency and maximum likelihood estimation probability (raw text using the lexicons).

Class	Count	% from neuter	% from total
n20	*17,065*	41.312%	9.081%
n20a	*23*	0.056%	0.012%
n21	*4,032*	9.761%	2.146%
n22	*5,655*	13.690%	3.009%
n23	*9,244*	22.378%	4.919%
n23a	*53*	0.128%	0.028%
n24	*1,757*	4.253%	0.935%
n25	*467*	1.131%	0.249%
n26	*608*	1.472%	0.324%
n27	*974*	2.358%	0.518%
n28	*215*	0.520%	0.114%
n28a	*124*	0.300%	0.066%
n29	*5*	0.012%	0.003%
n30	*1,037*	2.510%	0.552%
n31	*49*	0.119%	0.026%
Total	***41,308***	**100.00%**	**21.98%**

Table 19. Neuter: Class frequency and maximum likelihood estimation probability. Based on 8,5MB raw text using the lexicons.

Remark

We would like to note that it was possible (and a bit simpler) to estimate the class probability distributions from the Stem Lexicon. One has to be aware of this since the per-lexicon and the per raw text frequencies may differ a lot. Using per raw text distributions is much more reliable. We still use the lexicon but weight its entries according to their raw text frequencies.

Table 20, Table 21 and Table 22 show the corresponding probability distributions estimated from the Stem Lexicon. Although the distributions estimated both ways tend to follow the same general shape we can see some quite big differences. The class *m1* has almost twice less lexicon entries than *m4*, but has 45% more occurrences looking at the raw texts. Figure 27, Figure 28 and Figure 29 show comparison of the distributions in % for the genders taken separately. Figure 30 shows the sorted percents for all the 39 morphological classes.?

Class	Count	% from masculine	% from total
m1	*976*	19.922%	7.416%
m1a	*6*	0.122%	0.046%
m2	*644*	13.146%	4.893%
m3	*51*	1.041%	0.388%
m3a	*12*	0.245%	0.091%
m4	*1,939*	39.580%	14.733%
m5	*86*	1.755%	0.653%
m6	*328*	6.695%	2.492%
m7	*264*	5.389%	2.006%
m7a	*1*	0.020%	0.008%

m8	*392*	8.002%	2.978%
m9	*186*	3.797%	1.413%
m10	*7*	0.143%	0.053%
m11	*7*	0.143%	0.053%
Total	***4,899***	**100.00%**	**37.22%**

Table 20. Masculine: Class frequency and maximum likelihood estimation probability (Stem Lexicon).

Class	Count	% from feminine	% from total
f12	*3*	0.050%	0.023%
f13	*20*	0.331%	0.152%
f14	*116*	1.921%	0.881%
f14a	*4*	0.066%	0.030%
f15	*123*	2.037%	0.935%
f15a	*9*	0.149%	0.068%
f16	*2,671*	44.237%	20.295%
f17	*2,862*	47.400%	21.746%
f18	*229*	3.793%	1.740%
f19	*1*	0.017%	0.008%
Total	***6,038***	**100.00%**	**45.88%**

Table 21. Feminine: Class frequency and maximum likelihood estimation probability (Stem Lexicon).

Class	Count	% from neuter	% from total
n20	*843*	37.905%	6.405%
n20a	*1*	0.045%	0.008%
n21	*90*	4.047%	0.684%
n22	*192*	8.633%	1.459%
n23	*707*	31.790%	5.372%
n23a	*3*	0.135%	0.023%
n24	*289*	12.995%	2.196%
n25	*24*	1.079%	0.182%
n26	*1*	0.045%	0.008%
n27	*28*	1.259%	0.213%
n28	*28*	1.259%	0.213%
n28a	*4*	0.180%	0.030%
n29	*6*	0.270%	0.046%
n30	*1*	0.045%	0.008%
n31	*7*	0.315%	0.053%
Total	***2,224***	**100.00%**	**16.90%**

Table 22. Neuter: Class frequency and maximum likelihood estimation probability (Stem Lexicon).

Remark

One may ask why we think that the words with unknown stems are likely to follow the general properties of the known words. To answer this question it is important to explain the problem we deal with better. The purpose of the System is the identification and morphological classification of unknown words. Here unknown word means a word whose stem is *missing* from the Stem Lexicon and not a word that is *new* to German. The new words to the language are another problem. It is perfectly possible that some of the morphological classes are still active and able to accept new words (both *foreign* words and *neologisms*) while other can be less productive and even unable to do so. Thus, the per class distribution of the new words for German can differ a lot from the per class distribution of the words. Anyway, even the new words are much more likely to follow the general inflection rules of an existing morphological class rather than to follow a new paradigm.

We do not want to enter in more details here since this is out of our current scope. What is important to note is that the new words are only part of the words with unknown stems our System processes. Our current lexicon is limited and has about 13,000 stems. A lot of important common nouns are missing there (e.g. *das Wort*). Another important source of new words are the compounds. Unlike most other European languages, this is a very powerful process in German and our experiments show that 17,30% of all possible stems we generated for the words with unknown stems from the NEGRA corpus can be split as compounds. This perc?nt rises up to 38,20% of the unknown words if we look at the accepted stems only (see Table 23). These numbers are consistent with the results reported in (Adda-Decker M. and Adda G., 2000) where the compounds splitting resulted in out-of-vocabulary words (65k vocabulary was used) reduction from 5.2% to 4.2%. Thus, they achieved 19.24% reduction but using a larger vocabulary, different compounds identification and splitting strategy and what is much more important: they are interested in *all* out-of-vocabulary words while we are interested in *nouns only*. In fact the nouns are much more likely to produce compounds than other parts of speech are (the compound POS is determined by the POS of the last concatenated word).

The compounds are obtained as a result of the concatenation of known words and belong to the same morphological class as the last compound part. Unlike the inflexion and derivation, this is a very powerful process because it is generative and can theoretically produce an unlimited umount of words. The *inflection* and *derivation* are another potential sources of words with unknown stems. (Remember that some of the noun forms are known but their stem is unknown. It is also possible that we know a word but we do not know some of its inflected forms. In this the stem would be unknown as well.) Both can generate words with unknown stems from the *Stem Lexicon* view point but are not very poweful in generating really new words to German.

	All Possible Stems		**Accepted Stems**	
	count	**%**	**count**	**%**
Compounds	4,899	17.30%	4,800	38.20%
Ending Rules	6,563	23.17%	4,021	32.00%
Total Stems	**28,324**	**100.00%**	**12,567**	**100.00%**

Table 23. Unknown stems whose morphological class has been recognised through a compound or with an ending rule. (NEGRA corpus)

In fact the main source of new nouns in German are the *proper nouns* of persons, cities, companies etc. The other major source of new words (and to any other language) are the *foreign words*. Both proper nouns and foreign words currently represent a smaller portion of the unknown words compared to the compounds and the regular words missing from the lexicon. That is why we currently presuppose the words with unkown stem follow the same morphological properties as the known words do. It is only in case of a very large lexicon, which garantees that most of the words with uknown words are actually new to German (e.g. foreign and proper nouns), that it would be reasonable to study the new words derivation process and reestimate the probability distributions from Table 17, Table 18 and Table 19. ?

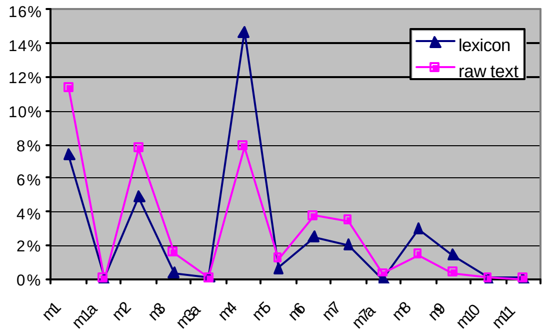


Figure 27. Masculine: Lexicon vs. raw text distribution (% from total).

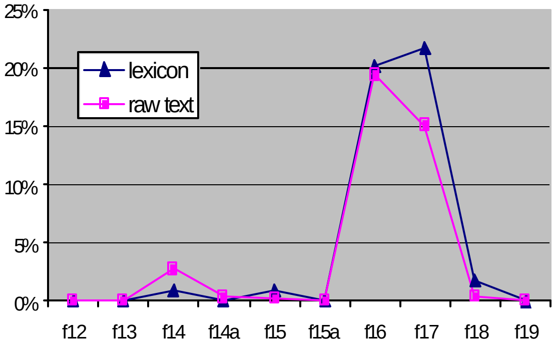


Figure 28. Feminine: Lexicon vs. raw text distribution (% from total).

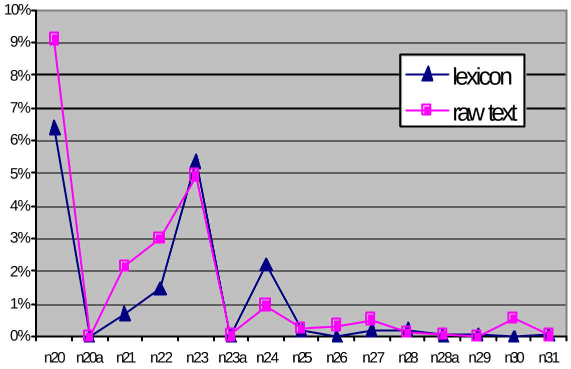


Figure 29. Neuter: Lexicon vs. raw text distribution (% from total).

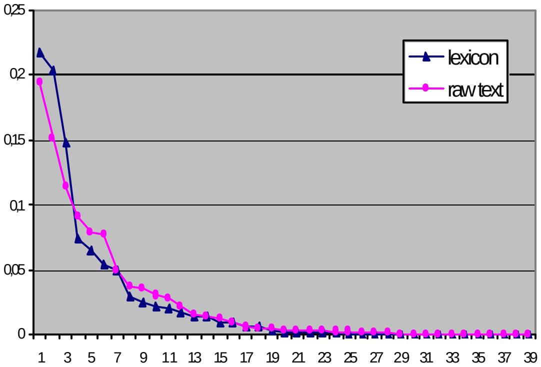


Figure 30. All, sorted in decreasing order: Lexicon vs. raw text distribution (% from total).

Let us now return to *Schrei* again. The first set {*Schrei*, *Schreien*} can be generated by the following classes {*m1*, *m8*, *m9*, *f12*, *f17*, *n20*, *n25*}, while {*Schrei*, *Schreier*} is compatible with {*m3a*, *n21*}. The most likely morphological class from the first set is *f17* (15.13%), while the one from the second class is *n21* (2.146%) and thus the first set wins (it is 7 times more likely!).

Another option is to compare the corresponding *sums* of probabilities for each set and not the better class. This is a better test since it checks more directly how likely is this combination. This time we get:

11.388% + 1.418% + 0.373% + 0.068% + 15.130% + 0.081% + 0.249% = **28.707%**

compared to

0.052% + 2.146% = **2.198%**

The first set wins again but this time it is 13 times more likely, while it was just 7 times more likely with the first test form. Although using the per-set probability sums may seem more reliable there is a problem with this approach. The main objection is that a word type set covered by more morphological classes will sum all their probabilities and probably win the test. But, as was mentioned above, it is unlikely that a stem is covered by more than one morphological class. In fact exactly one morphological class is likely to cover the word forms set given the right stem. The best-class test form implicitly accepts that this is the best class while the set-sum test considers all the classes. A careful evaluation is needed in order to decide what is better but for the moment we use the best-class strategy. This is a common issue and our tests show it happens about 10% of the time.

8.3.3 Algorithm

1. Go through the morphological classes and for each one:

   1.1. If the class covers more words than all the classes considered till now, save it.

1.2. If the class covers exactly the same amount of words than the best class till now and is more likely than the best one save it, otherwise — reject it

2. Keep the words covered by the best morphological class.

3. Go again through the morphological classes and find all that cover the words kept.

8.3.4 Demonstration

Table 24 and Table 25 demonstrate the algorithm at work. Table 24 lists the top unknown stems found in the NEGRA corpus ordered by word forms that generated the stem count and then alphabetically. There are some quite common words like *das Wort* and *der Ost*, which may seem strange, but their stems have not been generated during the automatic morphological classes induction process and thus are missing from the Stem Lexicon, which means they are unknown to the System. Table 25 shows the same list after stem refinements. Both lists contain all the stems covering at least 3 words. Thus, all stems that appeared in Table 24 but did not in Table 25 have been refined and cover two or one words after the refinement. We can see that some mechanically created stem groups like the one headed by *Bon*, which is supposed to cover the word types *Bona*, *Bonn* and *Bonus*, have been refined and disappeared from the second list (in fact they will appear below in the list but it has been cut at *Georg*). The stem *Bildungsurlaub*, which was initially supposed to cover 4 words was reduced to 3. The stem *West* lost the word form *Western* and thus was reduced to 3 word forms.

Unknown Stem	#	Words that *Generated* the Stem
Ortsbeirat	5	{ Ortsbeirat, Ortsbeirates, Ortsbeirats, Ortsbeiräte, Ortsbeiräten }
Bildungsurlaub	**4**	**{ Bildungsurlaub, Bildungsurlaube, Bildungsurlauben, Bildungsurlauber }**
Bo	4	{ Bo, Boer, Bose, Boses }
Gemeindehaushalt	4	{ Gemeindehaushalt, Gemeindehaushalte, Gemeindehaushaltes, Gemeindehaushalts }
Jo	4	{ Joe, Jon, Jos, Jose }
Kinderarzt	4	{ Kinderarzt, Kinderarztes, Kinderärzte, Kinderärzten }
Kunstwerk	4	{ Kunstwerk, Kunstwerke, Kunstwerken, Kunstwerks }
Lebensjahr	4	{ Lebensjahr, Lebensjahren, Lebensjahres, Lebensjahrs }
Ortsbezirk	4	{ Ortsbezirk, Ortsbezirke, Ortsbezirken, Ortsbezirks }
Ost	4	{ Ost, Osten, Oster, Ostern }
Stadtteil	4	{ Stadtteil, Stadtteile, Stadtteilen, Stadtteils }
West	**4**	**{ West, Weste, Westen, Western }**
Wort	4	{ Wort, Worte, Worten, Wortes }
Abend	3	{ Abend, Abende, Abenden }
Algerie	3	{ Algerien, Algeriens, Algerier }
Ander	3	{ Andere, Anderen, Anders }
Andre	3	{ Andrea, Andreas, Andres }
Anteilseigner	3	{ Anteilseigner, Anteilseignern, Anteilseigners }
Arbeitsplatz	3	{ Arbeitsplatz, Arbeitsplätze, Arbeitsplätzen }
Aufsichtsrat	3	{ Aufsichtsrat, Aufsichtsrates, Aufsichtsrats }
Augenblick	3	{ Augenblick, Augenblicken, Augenblicks }
Autofahr	3	{ Autofahren, Autofahrer, Autofahrern }
Band	3	{ Bandes, Bänder, Bändern }
Bau	3	{ Bau, Bauen, Baus }
Befreiungskampf	3	{ Befreiungskampf, Befreiungskampfes, Befreiungskämpfer }
Bensheim	3	{ Bensheim, Bensheimer, Bensheims }
Bernbach	3	{ Bernbach, Bernbacher, Bernbachs }
Biergarte	3	{ Biergarten, Biergartens, Biergärten }
Biergarten	3	{ Biergarten, Biergartens, Biergärten }
Bildungsurlaube	3	{ Bildungsurlaube, Bildungsurlauben, Bildungsurlauber }
Bon	**3**	**{ Bona, Bonn, Bonus }**
Brock	3	{ Brock, Brocks, Bröcker }
Bundesland	3	{ Bundesland, Bundesländer, Bundesländern }
Bürgerkrieg	3	{ Bürgerkrieg, Bürgerkrieges, Bürgerkriegs }
Edelstahlwerk	3	{ Edelstahlwerke, Edelstahlwerken, Edelstahlwerkes }
Edelstahlwerke	3	{ Edelstahlwerke, Edelstahlwerken, Edelstahlwerkes }
Eigentum	3	{ Eigentum, Eigentümer, Eigentümern }

Eigentümer	3	{ Eigentümer, Eigentümern, Eigentümers }
Energieplan	3	{ Energieplan, Energieplaner, Energieplans }
Erfolgsrezept	3	{ Erfolgsrezept, Erfolgsrezepten, Erfolgsrezepts }
Flörsheim	3	{ Flörsheim, Flörsheimer, Flörsheims }
Geist	3	{ Geist, Geiste, Geistes }
Georg	3	{ Georg, George, Georges }
Geschehen	3	{ Geschehen, Geschehene, Geschehens }
Grundrecht	3	{ Grundrecht, Grundrechte, Grundrechts }
Grundschul	3	{ Grundschule, Grundschulen, Grundschüler }
Gruppenspiel	3	{ Gruppenspiel, Gruppenspiele, Gruppenspielen }
Hanau	3	{ Hanau, Hanauer, Hanaus }
Herman	3	{ Herman, Hermann, Hermanns }
Hochmoor	3	{ Hochmoor, Hochmoore, Hochmooren }
Hundert	3	{ Hunderte, Hunderten, Hunderter }
Hunderte	3	{ Hunderte, Hunderten, Hunderter }
Idyll	3	{ Idylle, Idyllen, Idylls }
Indonesie	3	{ Indonesien, Indonesiens, Indonesier }
Ing	3	{ Ing, Inge, Inger }
Jugendzentr	3	{ Jugendzentren, Jugendzentrum, Jugendzentrums }
Karnevalverein	3	{ Karnevalverein, Karnevalvereine, Karnevalvereinen }
Kinderarzte	3	{ Kinderarztes, Kinderärzte, Kinderärzten }
Kindergarte	3	{ Kindergarten, Kindergartens, Kindergärten }
Kindergarten	3	{ Kindergarten, Kindergartens, Kindergärten }
Krankenhaus	3	{ Krankenhaus, Krankenhäuser, Krankenhäusern }
Kreisvorsitzend	3	{ Kreisvorsitzende, Kreisvorsitzenden, Kreisvorsitzender }
Kreisvorsitzende	3	{ Kreisvorsitzende, Kreisvorsitzenden, Kreisvorsitzender }
Langenhain	3	{ Langenhain, Langenhainer, Langenhains }
Lebenslauf	3	{ Lebenslauf, Lebenslaufes, Lebensläufe }
Leut	3	{ Leut, Leute, Leuten }
Mai	3	{ Mai, Maier, Main }
Munch	3	{ Munch, Munchs, München }
Musikzug	3	{ Musikzug, Musikzugs, Musikzüge }
Mörlenbach	3	{ Mörlenbach, Mörlenbachern, Mörlenbachs }
Name	3	{ Name, Namen, Namens }
Nicol	3	{ Nicola, Nicolas, Nicole }
Ortsbeirate	3	{ Ortsbeirates, Ortsbeiräte, Ortsbeiräten }
Papp	3	{ Papp, Pappe, Pappen }
Programmheft	3	{ Programmheft, Programmhefte, Programmheften }
Punkt	3	{ Punkt, Punkte, Punkten }
Regenwald	3	{ Regenwald, Regenwaldes, Regenwälder }
Sach	3	{ Sacher, Sachs, Sachsen }
Schmitt	3	{ Schmitt, Schmitten, Schmitts }
Schuldenberg	3	{ Schuldenberge, Schuldenberges, Schuldenbergs }
Sitzplatz	3	{ Sitzplatz, Sitzplätze, Sitzplätzen }
Spd-fraktionsvorsitzend	3	{ Spd-Fraktionsvorsitzende, Spd-Fraktionsvorsitzenden, Spd-Fraktionsvorsitzender }
Spd-fraktionsvorsitzende	3	{ Spd-Fraktionsvorsitzende, Spd-Fraktionsvorsitzenden, Spd-Fraktionsvorsitzender }
Spielplatz	3	{ Spielplatz, Spielplätze, Spielplätzen }
Spieltag	3	{ Spieltag, Spieltage, Spieltagen }
Sportplatz	3	{ Sportplatz, Sportplätze, Sportplätzen }
Sportverein	3	{ Sportverein, Sportvereine, Sportvereins }
Stadtteilparlament	3	{ Stadtteilparlament, Stadtteilparlamentes, Stadtteilparlaments }
Stadtverordnet	3	{ Stadtverordnete, Stadtverordneten, Stadtverordneter }
Stadtverordnete	3	{ Stadtverordnete, Stadtverordneten, Stadtverordneter }
Stahlwerk	3	{ Stahlwerk, Stahlwerke, Stahlwerker }
Stra?enbauamt	3	{ Stra?enbauamt, Stra?enbauamtes, Stra?enbauamts }
Sud	3	{ Sud, Süd, Süden }
Sv	3	{ Sv, Sva, Sven }
Tagebuch	3	{ Tagebuch, Tagebuchs, Tagebüchern }
Tarifvertrag	3	{ Tarifvertrag, Tarifvertrags, Tarifverträgen }
Tibet	3	{ Tibet, Tibeter, Tibetern }
Tod	3	{ Tod, Tode, Todes }
Vereinsheim	3	{ Vereinsheim, Vereinsheimen, Vereinsheims }
Verwaltungshaushalt	3	{ Verwaltungshaushalt, Verwaltungshaushaltes,

		Verwaltungshaushalts }
Worte	3	{ Worte, Worten, Wortes }
Zehntausend	3	{ Zehntausend, Zehntausende, Zehntausenden }

Table 24. Unknown stems: (NEGRA corpus) ordered by word types covered count.

Refined Unknown Stem	#	Words *Covered* by the stem
Ortsbeirat	5	{ Ortsbeirat, Ortsbeirates, Ortsbeirats, Ortsbeiräte, Ortsbeiräten }
Gemeindehaushalt	4	{ Gemeindehaushalt, Gemeindehaushalte, Gemeindehaushaltes, Gemeindehaushalts }
Kinderarzt	4	{ Kinderarzt, Kinderarztes, Kinderärzte, Kinderärzten }
Kunstwerk	4	{ Kunstwerk, Kunstwerke, Kunstwerken, Kunstwerks }
Lebensjahr	4	{ Lebensjahr, Lebensjahren, Lebensjahres, Lebensjahrs }
Ortsbezirk	4	{ Ortsbezirk, Ortsbezirke, Ortsbezirken, Ortsbezirks }
Stadtteil	4	{ Stadtteil, Stadtteile, Stadtteilen, Stadtteils }
Wort	4	{ Wort, Worte, Worten, Wortes }
Abend	3	{ Abend, Abende, Abenden }
Ander	3	{ Andere, Anderen, Anders }
Anteilseigner	3	{ Anteilseigner, Anteilseignern, Anteilseigners }
Arbeitsplatz	3	{ Arbeitsplatz, Arbeitsplätze, Arbeitsplätzen }
Aufsichtsrat	3	{ Aufsichtsrat, Aufsichtsrates, Aufsichtsrats }
Augenblick	3	{ Augenblick, Augenblicken, Augenblicks }
Band	3	{ Bandes, Bänder, Bändern }
Bau	3	{ Bau, Bauen, Baus }
Befreiungskampf	3	{ Befreiungskampf, Befreiungskampfes, Befreiungskämpfer }
Bensheim	3	{ Bensheim, Bensheimer, Bensheims }
Bernbach	3	{ Bernbach, Bernbacher, Bernbachs }
Biergarten	3	{ Biergarten, Biergartens, Biergärten }
Bildungsurlaub	**3**	**{ Bildungsurlaub, Bildungsurlaube, Bildungsurlauben }**
Bildungsurlaube	**3**	**{ Bildungsurlaube, Bildungsurlauben, Bildungsurlauber }**
Bo	3	{ Bo, Bose, Boses }
Brock	3	{ Brock, Brocks, Bröcker }
Bundesland	3	{ Bundesland, Bundesländer, Bundesländern }
Bürgerkrieg	3	{ Bürgerkrieg, Bürgerkrieges, Bürgerkriegs }
Edelstahlwerk	3	{ Edelstahlwerke, Edelstahlwerken, Edelstahlwerkes }
Edelstahlwerke	3	{ Edelstahlwerke, Edelstahlwerken, Edelstahlwerkes }
Eigentum	3	{ Eigentum, Eigentümer, Eigentümern }
Eigentümer	3	{ Eigentümer, Eigentümern, Eigentümers }
Energieplan	3	{ Energieplan, Energieplaner, Energieplans }
Erfolgsrezept	3	{ Erfolgsrezept, Erfolgsrezepten, Erfolgsrezepts }
Flörsheim	3	{ Flörsheim, Flörsheimer, Flörsheims }
Geist	3	{ Geist, Geiste, Geistes }
Georg	3	{ Georg, George, Georges }
Geschehen	3	{ Geschehen, Geschehene, Geschehens }
Grundrecht	3	{ Grundrecht, Grundrechte, Grundrechts }
Gruppenspiel	3	{ Gruppenspiel, Gruppenspiele, Gruppenspielen }
Hanau	3	{ Hanau, Hanauer, Hanaus }
Herman	3	{ Herman, Hermann, Hermanns }
Hochmoor	3	{ Hochmoor, Hochmoore, Hochmooren }
Hunderte	3	{ Hunderte, Hunderten, Hunderter }
Idyll	3	{ Idylle, Idyllen, Idylls }
Ing	3	{ Ing, Inge, Inger }
Jugendzentr	3	{ Jugendzentren, Jugendzentrum, Jugendzentrums }
Karnevalverein	3	{ Karnevalverein, Karnevalvereine, Karnevalvereinen }
Kinderarzte	3	{ Kinderarztes, Kinderärzte, Kinderärzten }
Kindergarten	3	{ Kindergarten, Kindergartens, Kindergärten }
Krankenhaus	3	{ Krankenhaus, Krankenhäuser, Krankenhäusern }
Kreisvorsitzende	3	{ Kreisvorsitzende, Kreisvorsitzenden, Kreisvorsitzender }
Langenhain	3	{ Langenhain, Langenhainer, Langenhains }
Lebenslauf	3	{ Lebenslauf, Lebenslaufes, Lebensläufe }
Leut	3	{ Leut, Leute, Leuten }
Munch	3	{ Munch, Munchs, München }
Musikzug	3	{ Musikzug, Musikzugs, Musikzüge }
Mörlenbach	3	{ Mörlenbach, Mörlenbachern, Mörlenbachs }

Name	3	{ Name, Namen, Namens }
Ortsbeirate	3	{ Ortsbeirates, Ortsbeiräte, Ortsbeiräten }
Ost	3	{ Ost, Oster, Ostern }
Papp	3	{ Papp, Pappe, Pappen }
Programmheft	3	{ Programmheft, Programmhefte, Programmheften }
Punkt	3	{ Punkt, Punkte, Punkten }
Regenwald	3	{ Regenwald, Regenwaldes, Regenwälder }
Schmitt	3	{ Schmitt, Schmitten, Schmitts }
Schuldenberg	3	{ Schuldenberge, Schuldenberges, Schuldenbergs }
Sitzplatz	3	{ Sitzplatz, Sitzplätze, Sitzplätzen }
Spd-fraktionsvorsitzende	3	{ Spd-Fraktionsvorsitzende, Spd-Fraktionsvorsitzenden, Spd-Fraktionsvorsitzender }
Spielplatz	3	{ Spielplatz, Spielplätze, Spielplätzen }
Spieltag	3	{ Spieltag, Spieltage, Spieltagen }
Sportplatz	3	{ Sportplatz, Sportplätze, Sportplätzen }
Sportverein	3	{ Sportverein, Sportvereine, Sportvereins }
Stadtteilparlament	3	{ Stadtteilparlament, Stadtteilparlamentes, Stadtteilparlaments }
Stadtverordnete	3	{ Stadtverordnete, Stadtverordneten, Stadtverordneter }
Stahlwerk	3	{ Stahlwerk, Stahlwerke, Stahlwerker }
Stra?enbauamt	3	{ Stra?enbauamt, Stra?enbauamtes, Stra?enbauamts }
Tagebuch	3	{ Tagebuch, Tagebuchs, Tagebüchern }
Tarifvertrag	3	{ Tarifvertrag, Tarifvertrags, Tarifverträgen }
Tibet	3	{ Tibet, Tibeter, Tibetern }
Tod	3	{ Tod, Tode, Todes }
Vereinsheim	3	{ Vereinsheim, Vereinsheimen, Vereinsheims }
Verwaltungshaushalt	3	{ Verwaltungshaushalt, Verwaltungshaushaltes, Verwaltungshaushalts }
West	**3**	**{ West, Weste, Westen }**
Worte	3	{ Worte, Worten, Wortes }
Zehntausend	3	{ Zehntausend, Zehntausende, Zehntausenden }

Table 25. ***Refined*** **unknown stems:** (NEGRA corpus) ordered by word types covered count.

8.4 Morphological stem analysis

Each stem generated in the previous step is analysed morphologically in order to obtain some additional information that could imply useful constraints on the subsequent analysis. The idea behind is that the more consistent knowledge we have about a stem the more likely it is to be the true stem for the word types it covers. The morphological analysis is based on both lexicon-based and suffix-based morphology.

- ***Lexicon-based morphology***
  - *Checking against the Stem Lexicon*
  - *Compounds splitting*
- ***Suffix-based morphology***
- ***Ending-based morphology***

8.4.1 Lexicon based morphology

8.4.1.1 Checking against the Stem Lexicon

We use the Stem Lexicon to check the unknown stems validity. In case a stem is found in the Stem Lexicon, we reject it. This is because of the assumption that all the stems in the Stem Lexicon are well-known (we know their morphological class). Thus, we force all their inflexions to be present in the Expanded Stem Lexicon. This means that no word type with unknown stem could have a known stem since all words a known stem generates are known.

8.4.1.2 Compounds splitting

An interesting problem are the German compound nouns. The concatenation of words is very common in German and it is not trivial to solve. These can contain base forms as well as inflected ones, e.g. *Haus-meister* but *Häuser-meer*. These can also be ambiguous: *Stau-becken* vs. *Staub-ecken*. The letters *e*, *s* and *n* can appear in the middle of a compound word: *Schwein-e-bauch*, *Schwein-s-blas*, but it is not strictly necessary: *Schwein-kram*. Anyway, for our algorithm none of these can be a problem since we simply try all the splits and if there is an *s*, an *e* or an *n* we try to remove it. In case an ambiguous splitting occurs we keep all the possible classes and leave the disambiguation for the subsequent steps. Special care is taken about the three-consonant rule.

Another interesting approach is by means of longest matching substrings found in the lexicon. Thus, a word like *adfadfeimer* will return as a result *eimer* assuming that *adfadf* is no legal lexical stem. (Neumann and Mazzini, 1999; Neumann et al., 1997)

(Adda-Decker & Adda, 2000) propose and test several different approaches including general rules for morpheme boundary identification. These are hypothesised after the occurrence of sequences such as: *-ungs*, *-hafts*, *-lings*, *-tions*, *-heits*.

Remark

It is important to note as well that this operation is highly lexicon dependent. Suppose that our lexicon contains the word *Staub* but not *Stau*. Thus, we will discover the reading *Staub-ecken*, which is very unlikely, and will miss the much more acceptable *Stau-becken*.

Our Stem Lexicon contains both *der Stau* (*m1*) and *das Becken* (*n23*) and at the same time it contains neither *der Staub* nor *die Ecken*. Thus, it will permit us to reveal the *Stau-becken* reading only. Although this is the correct reading we just had chance and sometimes we will fail. ?

ab
an
auf
aus
auseinander
bei
da
dar
davon
durch
ein
empor
entgegen
fertig
fest
fort
frei
heim
her
herab
heran
herauf
heraus
herbei
heruber
herum
herunter
hervor
hier
hierher
hinab
hinauf
hinaus
hinein
los
mit
nach
nieder
satt
teil
ubel
um
umher
unter
voll
vor
voran
voraus
vorwarts
wahr
weg
weiter
wieder
wiederher
zu
zurecht
zuruck
zusammen

Figure 31. Separable prefixes we look for when splitting compounds.

There is a second way the lexicon may be used. If we add or remove a standard grammatical German prefix to the unknown stem and this generates a known stem, then we think they must belong to the same morphological class and we thus attach the class of the known stem to the unknown one. We cannot do this with suffixes since the known stem thus obtained will not provide any information about the unknown one.

In fact we currently do not try to add prefixes since it is not very likely that our Stem Lexicon contains a stem form with a prefix and at the same time does not contain the form without prefix, although this is possible. For the moment we prefer to keep trying prefix removal but not addition since this may introduce errors. The prefix removal currently is integrated in the compound splitting algorithm. We are looking for separable prefixes at the beginning of a compound word, see Figure 31.

Before we explain the compound splitting in more details we need some definitions.

Definition 1:

A character string is a *legal compound member*, iff it is present in the Word Lexicon and has one of the following tags: *ADJ*, *ADV*, *PRO*, *ART*, *PA1*, *PA2*, *PRP*, *SUB*, *VER*, *ZAL*.

Definition 2:

A character string *s* is an *acceptable compound beginning* if exactly one of the following holds:

1) It is a separable prefix from a pre-specified list
2) It is a legal compound member.
3) It can be split in two strings *a* and *b*, such that

- *s* = *ab*
- *a* is an acceptable compound beginning
- *b* is a legal compound member

How we split the compounds? Given a stem we go from right to left, cut its last few characters and check whether they represent a known stem. If so, we check whether the remaining first part is an *acceptable compound beginning*. Note that, while the correct stem identification is very important and we try to find all the possibilities, we are much more liberal in what about the first part of the compound. In case the first part is composed of more than one word it could be possible to split it in more than way. This time this is unimportant and we are interested just in whether this is possible or not and not how exactly this may be done. What really matters is the second part because it determines the morphological class of the whole compound. We check the first part just to be sure this is really a compound. If we do not check it we could introduce errors. Consider for example the stem *Direktor* and suppose it is unknown. One may try to split it as *Direk-tor* and if we do not check the first part we get the stem *Tor*, which is present in the Stem Lexicon with the class *m8*. Using this split leads to an error since *der Direktor* has class *m9*. But if we try to check whether the first part *Direk* is an *acceptable compound beginning* we will see it is not (*Direk* is missing from the *Word Lexicon* and in plus cannot be split in a way that will permit us to say it is an acceptable compound beginning) and thus reject this splitting.

8.4.2 Dictionary-based suffix morphology

Another source of information we could exploit are some regularities in German regarding the stem suffixes. Some of the suffixes are highly predictive and can indicate the morphological class or just the gender. (We cannot expect a stem suffix to show features like case or number since they are a property of the *inflected form* and have nothing to do with the stem suffix). Our tests show that usually, if an ending is a good predictor for the gender, it is a good predictor for some morphological class as well (see Table 29, Table 30 and Table 31).

The German grammar (*Drosdowski G., 1984*) provides a list of some characteristic suffixes revealing the noun gender. Some of these are ambiguous and may have exceptions. Some of the exceptions are listed in the grammar but the list is not exhaustive.

Masculine Suffix	Examples	Exceptions
ich	der Teppich	
ig	der Honig	*das Reisig*
ling	der Fremdlings	*die Reling*
s	der Schnaps	
and	der Doctorand	
ant	der Aspirant	
är	der Militär	*das Salär*
ast	der Dynast	
eur	der Amateur	
ör	der Likör	
[i]ent	der Skribent, der Interessent	
ier	der Bankier	*das Kollier,das Spalier,die Manier*
iker	der Graphiker	
ikus ismus ist or	der Musikus der Realismus der der Pianist der Motor	

Table 26. Masculine suffixes (German grammar).

Feminine Suffix	Examples	Exceptions
ei	die Reiberei	
in	die Freundin	
heit	die Einheit	
keit	die Kleinigkeit	
schaft	die Herrschaft	
ung	die Achtung	*der Hornung*
a	die Kamera	
ade	die Kanonade	
age	die Garage	
aille aise	die Bataille die Marseillaise	
äse	die Polonäse	
ance	die Renaissance	
äne	die Fontäne	
anz	die Arroganz	
ation	die Oxydation	
elle	die Morelle	
ette	die Tolette	
euse ie	die Friseuse die Materie	 *das Genie*
[i]enz	die Audienz, die Prominenz	
[i]ere	die Voliere, die Misere	
ik	die Musik	
ille	die Bastille	
ine	die Kabine	
ion	die Explosion	
isse	die Mantisse	
[i]tät itis	die Vitalität die Rachitis	
ive	die Direktive	
ose	die Neurose	
se	die Base	
sis	die Basis	
ur	die Natur	
üre	die Bordüre	

Table 27. Feminine suffixes (German grammar).

Neuter Suffix	Examples	Exceptions
chen	das Mädchen	
lein	das Ingelein	
le	das Mariele	
icht	das Dickicht	
tel,teil	das Viertel	
tum	das Volkstum	*der Irrtum, der Reichtum*
ett	das Balett	
in	das Benzin	
[i]um	das Album	
ma	das Komma	
ment	das Dokument	

Table 28. Neuter suffixes (German grammar).

Masculine Suffixes

ich

37 m1
3 n20

ig

1 f15
2 f15a
10 m1
1 n24

ling

1 f15
38 m1
1 m6

s

19 f13
11 f14
1 f16
1 f17
32 m1
6 m1a
10 m2
1 m8
1 m9
10 n20
10 n22
28 n27

and

13 f14
1 f15
1 m1
38 m2
2 m3

7	m8
16	n22

ant	
3	m6
44	m8
1	n24
1	n25

är	
13	m1
3	m8

ast	
7	f17
3	m1
4	m2
1	m6
3	m8
4	m9

eur	
12	m1

ör	
2	m1
1	n20
1	n25

ent	
6	m1
39	m8
34	n20
1	n21
5	n24

ient	
7	m8

ier	
5	f16
1	f17
14	m1
25	m4
2	m6
39	n20
2	n24

iker	
74	m4

ikus	
%	

ismus	
2	m11

ist	
5	f17
2	m1
6	m3a
138	m8

or	
3	m1
2	m2
1	m4
1	m6
4	m8
153	m9
5	n20
2	n24

Table 29. Masculine endings distributions (Stem Lexicon).

Feminine Suffixes

ei	
1	f16
166	f17
4	m1
1	m9
1	n21
1	n24

in	
2	f17
229	f18
41	m1
1	m4
4	m6
35	n20
72	n23
7	n24

heit	
70	f17
1	m1

keit	
137	f17

schaft	
65	f17

ung	
1443	f17
16	m2
1	m8

a	
57	f15
9	f15a
26	m6
20	n24
4	n28a

ade	
36	f16
1	m7
1	n23

age	
85	f16
3	n23

aille	
2	f16

aise	
5	f16

äse	
1	f16
5	m4
1	n23

ance	
4	f16

äne	
4	f16
2	m7

anz	
22	f17
3	m2

ation	
1	f15

229	f17

elle	
80	f16
1	m7

ette	
43	f16
2	n24

euse	
1	f16

ie	
1	f15
197	f16
7	m6
2	m7
1	n23
2	n24

enz	
43	f17

ienz	
2	f17

ere	
21	f16

iere	
9	f16

ik	
36	f17
2	m6
2	m8

ille	
15	f16

ine	
2	f15
70	f16
1	m7
1	n24

ion	
1	f15
433	f17
2	m1

isse	
1	f13
7	f16

tät	
57	f17

ität	
55	f17

itis	
%%%	

ive	
12	f16
1	m6

ose	
39	f16
4	m7

se	
1	f13
256	f16
5	m4
28	m7
6	n23

sis	
%%%	

ur	
1	f12
1	f14
72	f17
13	m1
1	m2
1	n20

üre	
9	f16

Table 30. Feminine endings distributions (Stem Lexicon).

Neuter Suffixes

chen	
12	m4
256	n23

lein	
72	n23
1	n24

le	
2	f15
228	f16
1	m10
1	m6
9	m7
4	n23
2	n24

icht	
19	f17
10	m1
23	n20
17	n21

tel	
6	f16
23	m4
2	m5
46	n23
2	n24

teil	
10	m1
13	n20

tum	
3	m3
5	n22
2	n28
2	n29

ett	
21	n20
7	n21
4	n24
5	n25

in	
2	f17
229	f18
41	m1
1	m4
4	m6
35	n20
72	n23
7	n24

ium	
18	n28

um	
27	m2
3	m3
1	m6
5	n22
2	n24
28	n28
6	n29

ma	
6	f15
1	f15a
3	m6
7	n24
4	n28a

ment	
3	m1
1	m8
30	n20
1	n21
5	n24

Table 31. Neuter endings distributions (Stem Lexicon).

The results we obtained may look quite strange at first glance: a lot of the suffixes are ambiguous. And what is stranger is that for most of them the grammar does not provide any exception. So, then the grammar is simply incorrect? The grammar provides a list of highly predictive *suffixes* while we look at *endings*. The grammar suggested suffixes are morphologically motivated while the results in Table 29, Table 30 and Table 31 are obtained looking at the endings without taking care whether the ending is really a suffix. Another source of ambiguity is a result of ending intersection. Looking at Table 26 we discover the masculine suffixes *-ikus*, *-ismus* and *-s*. Table 27 contains the suffixes *-itis* and *-is*. All these end on *-s* and thus have been counted under the ending *–s* in Table 29. (In fact the Stem Lexicon does contain words ending on neither *–ikus* nor *–sis*). Another strange example is the suffix *–in*, which is listed as both feminine (see Table 27) and neuter (see Table 28). Not suprisingly, both the feminine (231 stems) and neuter (114 stems) examples are met in the Stem Lexicon. What is interesting is that there are 46 masculine stems. Which is neither predicted by the dictionary as a rule (*-in* is not listed as masculine ending, see Table 26) nor is listed as exception. And unlike the case with *–s* this time there are no masculine endings that could be mixed with *–in*.

8.4.3 Probabilistic ending guessing morphology

While most of the rules generated through the dictionary suggested suffix morphology seem to be good predictors for either gender or morphological class the failures of the method made us think of more systematic alternative way for automatic ending guessing rules generation. We implemented a Mikheev-style ending guessing rules (Mikheev, 1997). He originally made this for POS guessing but we applied the same approach for morphological class guessing. We selected confidence level of 90% and considered endings up to 7 characters long that must be preceded by at least 3 characters. We did this once against the Stem Lexicon and then against a raw text by checking the words against the Expanded Stem Lexicon and from there against the Stems Lexicon. We keep only the rules with confidence score at least 0.90 and frequency at least 10. This resulted in 482 rules when running the rules induction against the Stem Lexicon and in 1789 rules when the Stem Lexicon entries were weighted according to their frequencies in a 8,5 MB raw text.

8.4.3.1 Ending guessing rules induction algorithm

We consider all the endings that are up to 7 characters long. Table 27 shows some of the German suffixes are up to 6 characters long, e.g. the highly predictive suffix *-schaft* (see Table 30). Thus, it is necessary to consider endings at least 6 characters long. We added one character more. We would like to stress again that the rules we are trying to induce are *ending* guessing rules and not *suffix* guessing rules. The automatic suffix identification is a hard task. In plus using suffixes only may prevent us from identifying some highly predictive endings. Previous research on automatic POS ending guessing rules induction have shown that some of the highest quality predictive endings are not standard suffixes. In plus, the results above (see Table 29, Table 30, Table 31) show that the standard suffixes may be of a particularly bad quality although listed in the grammar as good predictors.

We consider *all* the endings up to 7 characters long that are met at least 10 times in the *training text* (we will explain the notion of training text below). For each noun token we extract all its endings. We consider the last *k* (*k*=1,2,...,7) characters represent a word ending if there removal leaves at least 3 characters including at least one vowel (does not matter whether short or long). For each rule we collect list of the morphological classes it appeared with together with the corresponding frequencies. We would like to accept as ending guessing rules only the top highly predictive ones. It is intuitively clear that a good ending guessing rule is:

- **unambiguous**

It predicts a particular class without or with only few exceptions. The fewer the exceptions the better the rule.

- **frequent**
  The rule must be based on large number of occurrences. The higher the occurrence number the more confident we are in the rule’s prediction and the higher the probability that an unknown stem will match it.
- **long**
  The ending length is another important consideration. The longer the rule the less the probability that it will appear due to chance and thus the better its prediction.

What we need is a score for the rules that takes into account at least these three concerns (and possibly more). Undoubtenly, the most important factor is the rule ambiguity. We would like our rule to be as accurate as possible with only few exceptions. A good predictor of the rule accuracy is the *maximum likelihood estimation* given by the formula:

$$\hat{p} = \frac{x}{n}$$

where:

x — the number of successful rule guesses
n — the total training stems compatible with the rule

Given a large set of training words we can find x_i and n_i for each ending guessing rule-candidate i. One way to do so is to investigate the stems from the Stem Lexicon: we count the stems n that are compatible with the rule and those of them whose morphological class has been correctly predicted by the rule: x. This is not a very good idea since the words the stems represent are not equally likely to be met in a real text. It is much better to estimate the frequencies x and n from a large collection of raw text. In this case we consider the words whose stem is known (the ones from the Expanded Stem Lexicon). This time the count n is the sum of the frequencies of all words whose stem is known and is compatible with the rule. The count x is estimated the same way from the raw text words whose morphological class has been correctly predicted by the rule.

Although the maximum likelihood estimation is a good predictor it does take into account neither the rule length nor the rule frequency. Thus, a rule that has just one occurrence in the corpus and has a correct prediction will receive the maximum score 1. A rule with 1000 occurrences, all of which have been correctly classified, will receive the same score. This is not what we would like to obtain since in the first case the correct prediction may be due just to *chance* while in the second case this is 1000 times less likely. In plus, as has been mentioned above, a more frequent rule is better since it is expected to cover more unknown stems than a less frequent one. Of course this depends a lot on the raw text used during the training. It must be as representative of the real language as possible. Usually, a large text collection is used mostly from newspapers since they are supposed to be very representative of the contemporary language and to cover a large amount of different fields.

So, what we saw above is that although the maximum likelihood estimation is a good predictor of the rule accuracy it is a bad predictor of the practical rule efficiency, which is mostly due to the insufficient amount of occurrences observed. (Mikheev, 1997) proposes a good solution to the problem. He substitutes the maximum likelihood estimation with the *minimum confidence limit* p, which gives the minimum expected value of $\hat{p}$ in case a large number of experiments have been performed. The minimum confidence level is given by the following formula:

$$p = p - t_{(1-a)/2}^{(n-1)} \sqrt{\frac{p(1-p)}{n}}$$

where

p is a modified version of $\hat{p}$ that ensures neither p nor $(1–p)$ can be zero: $p = (x+0.5)/(n+1)$

$\sqrt{\frac{p(1-p)}{n}}$ is an estimation of the dispersion

$t^{(n-1)}_{(1-a)/2}$ is a coefficient of the *t*-distribution.

The *t*-distribution $t^{d}_{(1-a)/2}$ has two parameters: the degree of freedom *d* and the confidence level a. Table 32 shows the values of the *t*-statistic for the confidence level of 0.900, 0.950, 0.975, 0.990 and 0.999. Mikheev suggests 0.90 degree of confidence and this is the level we use currently.

The minimum confidence limit is a better predictor of the rule quality and takes into account the rule frequency. But it still does not prefer longer rules to shorter ones other things being equal. (Mikheev, 1997) proposes to use the logarithm of the ending length *l* in a score of the form:

$$score = p - \frac{t^{(n-1)}_{(1-a)/2}\sqrt{\dfrac{p(1-p)}{n}}}{1+\log(l)}, p = (x+0.5)/(n+1)$$

This is the final form of the score calculation formula proposed by Mikheev. It is easy to see that the score values are between 0 and 1. He scores all the rules that are met at least twice and selects only the ones above a certain threshold. (Mikheev, 1997) suggests thresholds in the interval 0.65-0.80 points but we use 0.90 in order to obtain rules of higher quality (but less in number). (Mikheev, 1997).

Degree of freedom	0.900	0.950	0.975	0.990	0.999
1	3.078	6.314	12.706	31.821	318.313
2	1.886	2.920	4.303	6.965	22.327
3	1.638	2.353	3.182	4.541	10.215
4	1.533	2.132	2.776	3.747	7.173
5	1.476	2.015	2.571	3.365	5.893
6	1.440	1.943	2.447	3.143	5.208
7	1.415	1.895	2.365	2.998	4.782
8	1.397	1.860	2.306	2.896	4.499
9	1.383	1.833	2.262	2.821	4.296
10	1.372	1.812	2.228	2.764	4.143
11	1.363	1.796	2.201	2.718	4.024
12	1.356	1.782	2.179	2.681	3.929
13	1.350	1.771	2.160	2.650	3.852
14	1.345	1.761	2.145	2.624	3.787
15	1.341	1.753	2.131	2.602	3.733
16	1.337	1.746	2.120	2.583	3.686
17	1.333	1.740	2.110	2.567	3.646
18	1.330	1.734	2.101	2.552	3.610
19	1.328	1.729	2.093	2.539	3.579
20	1.325	1.725	2.086	2.528	3.552
21	1.323	1.721	2.080	2.518	3.527
22	1.321	1.717	2.074	2.508	3.505
23	1.319	1.714	2.069	2.500	3.485
24	1.318	1.711	2.064	2.492	3.467
25	1.316	1.708	2.060	2.485	3.450
26	1.315	1.706	2.056	2.479	3.435
27	1.314	1.703	2.052	2.473	3.421
28	1.313	1.701	2.048	2.467	3.408
29	1.311	1.699	2.045	2.462	3.396
30	1.310	1.697	2.042	2.457	3.385
31	1.309	1.696	2.040	2.453	3.375
32	1.309	1.694	2.037	2.449	3.365
33	1.308	1.692	2.035	2.445	3.356

34	1.307	1.691	2.032	2.441	3.348
35	1.306	1.690	2.030	2.438	3.340
36	1.306	1.688	2.028	2.434	3.333
37	1.305	1.687	2.026	2.431	3.326
38	1.304	1.686	2.024	2.429	3.319
39	1.304	1.685	2.023	2.426	3.313
40	1.303	1.684	2.021	2.423	3.307
41	1.303	1.683	2.020	2.421	3.301
42	1.302	1.682	2.018	2.418	3.296
43	1.302	1.681	2.017	2.416	3.291
44	1.301	1.680	2.015	2.414	3.286
45	1.301	1.679	2.014	2.412	3.281
46	1.300	1.679	2.013	2.410	3.277
47	1.300	1.678	2.012	2.408	3.273
48	1.299	1.677	2.011	2.407	3.269
49	1.299	1.677	2.010	2.405	3.265
50	1.299	1.676	2.009	2.403	3.261
51	1.298	1.675	2.008	2.402	3.258
52	1.298	1.675	2.007	2.400	3.255
53	1.298	1.674	2.006	2.399	3.251
54	1.297	1.674	2.005	2.397	3.248
55	1.297	1.673	2.004	2.396	3.245
56	1.297	1.673	2.003	2.395	3.242
57	1.297	1.672	2.002	2.394	3.239
58	1.296	1.672	2.002	2.392	3.237
59	1.296	1.671	2.001	2.391	3.234
60	1.296	1.671	2.000	2.390	3.232
61	1.296	1.670	2.000	2.389	3.229
62	1.295	1.670	1.999	2.388	3.227
63	1.295	1.669	1.998	2.387	3.225
64	1.295	1.669	1.998	2.386	3.223
65	1.295	1.669	1.997	2.385	3.220
66	1.295	1.668	1.997	2.384	3.218
67	1.294	1.668	1.996	2.383	3.216
68	1.294	1.668	1.995	2.382	3.214
69	1.294	1.667	1.995	2.382	3.213
70	1.294	1.667	1.994	2.381	3.211
71	1.294	1.667	1.994	2.380	3.209
72	1.293	1.666	1.993	2.379	3.207
73	1.293	1.666	1.993	2.379	3.206
74	1.293	1.666	1.993	2.378	3.204
75	1.293	1.665	1.992	2.377	3.202
76	1.293	1.665	1.992	2.376	3.201
77	1.293	1.665	1.991	2.376	3.199
78	1.292	1.665	1.991	2.375	3.198
79	1.292	1.664	1.990	2.374	3.197
80	1.292	1.664	1.990	2.374	3.195
81	1.292	1.664	1.990	2.373	3.194
82	1.292	1.664	1.989	2.373	3.193
83	1.292	1.663	1.989	2.372	3.191
84	1.292	1.663	1.989	2.372	3.190
85	1.292	1.663	1.988	2.371	3.189
86	1.291	1.663	1.988	2.370	3.188
87	1.291	1.663	1.988	2.370	3.187
88	1.291	1.662	1.987	2.369	3.185

89	1.291	1.662	1.987	2.369	3.184
90	1.291	1.662	1.987	2.368	3.183
91	1.291	1.662	1.986	2.368	3.182
92	1.291	1.662	1.986	2.368	3.181
93	1.291	1.661	1.986	2.367	3.180
94	1.291	1.661	1.986	2.367	3.179
95	1.291	1.661	1.985	2.366	3.178
96	1.290	1.661	1.985	2.366	3.177
97	1.290	1.661	1.985	2.365	3.176
98	1.290	1.661	1.984	2.365	3.175
99	1.290	1.660	1.984	2.365	3.175
100	1.290	1.660	1.984	2.364	3.174

Table 32. Values of the *t*–statistics.

8.4.3.2 Stem Lexicon estimation

The ending guessing have been estimated twice: once directly from the Stem Lexicon and once from a raw text collection. Table 33 and Table 34 show the top and the bottom part (the rules just above the threshold of 0.90) of the ranked rules list.

Ending	Confidence	Class(es)	Frequency
erung	0.997051	f17	288
eit	0.996159	f17	247
tung	0.995234	f17	186
ler	0.995005	m4	190
ierung	0.994828	f17	159
tion	0.99396	f15 f17	1 358
gung	0.993809	f17	143
keit	0.993632	f17	139
ion	0.992006	m1 f15 f17	1 1 436
dung	0.991739	f17	107
nung	0.991421	f17	103
ation	0.990751	f15 f17	1 226
igkeit	0.99001	f17	82
gkeit	0.989818	f17	83
cher	0.989734	m4	86
rer	0.989256	m4	88
hung	0.989236	f17	82
igung	0.98917	f17	78
schaft	0.987968	f17	68
ung	0.987709	m2 m8 f17	14 1 1448
chaft	0.987589	f17	68
heit	0.98722	f17	69
iker	0.98722	m4	69
haft	0.987035	f17	68
lung	0.986651	m8 f17	1 161
chung	0.985467	f17	58
sion	0.985071	f17	59
ndung	0.9844	f17	54
ktion	0.9844	f17	54
ität	0.983994	f17	55
tät	0.983454	f17	57
zung	0.982751	f17	51
ner	0.982565	m4 n23	129 1

bung	0.982407	f17	50
ichkeit	0.98231	f17	45
rin	0.982016	f18 n20	125 1
chkeit	0.981887	f17	45
zer	0.981879	m4	52
sung	0.981682	f17	48
sierung	0.981496	f17	43
ker	0.981431	m4 n23	121 1
hkeit	0.981313	f17	45
htung	0.980892	f17	44
kung	0.980474	f17	45
llung	0.980452	f17	43
erei	0.980395	f16 f17	1 109
lle	0.979958	m7 f16	1 112
erin	0.979663	f18 n20	105 1
chtung	0.978592	f17	38
fung	0.978587	f17	41
enz	0.978119	f17	43
tation	0.977415	f17	36
nist	0.976917	m8	38
ration	0.97678	f17	35
ellung	0.97678	f17	35

Table 33. Stem Lexicon: Best-quality Mikheev-style ending guessing rules, confidence level 90%.

Ending	Confidence	Class(es)	Frequency
adt	0.9161	f14	11
ive	0.9161	f16	11
mie	0.9161	f16	11
ramm	0.915712	m1 n20	1 24
unde	0.915712	m7 f16	1 24
?e	0.915579	m7 f16	1 28
ist	0.915472	m1 m3a m8 f17	1 6 137 3
elung	0.914988	m8 f17	1 23
che	0.91477	m7 f16 f19	7 112 1
beit	0.914299	f17	10
nitt	0.914299	m1	10
ille	0.914299	f16	10
hein	0.914299	m1	10
dner	0.914299	m4	10
fach	0.914299	n22	10
dt	0.911686	f14	12
amm	0.911442	m1 n20	1 24
punkt	0.911412	m1 m9	22 1
pf	0.909363	m1 m2	2 39
atte	0.908612	m7 f16	1 22
unkt	0.908612	m1 m9	22 1
tin	0.908177	f18 n20	47 2

		n24	1
nkt	0.907857	m1	23
		m9	1
ieb	0.907837	m1	10
ruf	0.907837	m1	10
ng	0.906427	m1	44
		m2	70
		m3a	1
		m6	6
		m8	1
		f15	2
		f17	1450
		n20	2
		n24	13
sche	0.905858	m7	2
		f16	44
		f19	1
le	0.905319	m6	1
		m7	9
		m10	1
		f15	2
		f16	223
		n23	4
		n24	2
oge	0.905082	m7	34
		m10	1
		f16	1
ahl	0.903974	m9	1
		f17	22
eb	0.903747	m1	11
öl	0.903747	n20	11
änder	0.903256	m4	20
		n23	1
ich	0.902447	m1	33
		n20	2

Table 34. Stem Lexicon: Bottom Mikheev-style ending guessing rules, confidence level 90%.

8.4.3.3 Raw text estimation

We estimated the endings scores again, this time from a raw text. Table 35 and Table 36 show the top and the bottom part (the rules just above the threshold of 0.90) of the ranked rules list.

Ending	Confidence	Class(es)	Frequency
heit	0.999496	f17	1761
nung	0.999458	f17	1638
schaft	0.999427	f17	1439
keit	0.999412	f17	1510
chaft	0.999409	f17	1439
tung	0.999408	f17	1498
gung	0.999394	f17	1464
haft	0.999383	f17	1439
lung	0.999182	f17	1084
nheit	0.999118	f17	964
tand	0.999066	m2	950
erung	0.999025	f17	872
dung	0.99894	f17	837
enheit	0.998938	f17	776
ger	0.998864	m4	836
gkeit	0.998777	f17	695
hte	0.998771	f16	773
igkeit	0.998768	f17	669
tion	0.998714	f17	690
lschaft	0.99871	f17	624
ling	0.998705	m1	685
itt	0.99867	m1	714

ler	0.998664	m4	711
genheit	0.998565	f17	561
ritt	0.998536	m1	606
ichkeit	0.998419	f17	509
chkeit	0.998382	f17	509
tag	0.998343	m1	573
htung	0.998335	f17	510
hkeit	0.998331	f17	509
hung	0.998275	f17	514
llung	0.99824	f17	483
eit	0.998174	m1 f17	5 3822
len	0.998139	m4	510
ndung	0.998131	f17	455
ion	0.998114	m1 f17	1 1207
egung	0.998111	f17	450
tling	0.998032	m1	432
u?	0.998027	m2	548
ptling	0.997985	m1	409
indung	0.997848	f17	383
igung	0.997837	f17	393
ache	0.997797	f16	403
kung	0.997741	f17	393
hältnis	0.99772	n27	353
talt	0.997689	f17	384
ältnis	0.997666	n27	353
ltnis	0.997593	n27	353
nschaft	0.997561	f17	330
dt	0.997551	f14	442
tellung	0.997539	f17	327
präch	0.997537	n20	345
ation	0.99753	f17	344
chte	0.997528	f16	359
atz	0.997513	m2	382
chichte	0.997509	f16	323
ellung	0.997481	f17	327
ühl	0.99748	n20	377
hichte	0.99745	f16	323
räch	0.997428	n20	345
ichte	0.997426	f16	330
druck	0.997418	m1	329
gang	0.997406	m2	342
ianer	0.997354	m4	321
aner	0.997351	m4	335
rheit	0.997251	f17	309
äch	0.997247	n20	345
stand	0.997242	m2	308
hste	0.997092	m7	305
chung	0.997092	f17	292
chtung	0.997059	f17	280
lo?	0.997005	n22	317
anken	0.996989	m4	282
spiel	0.996989	n20	282
fnung	0.996989	f17	282
hrung	0.996957	f17	279
eger	0.996931	m4	289
lu?	0.996907	m2	307
piel	0.996877	n20	284

Table 35. Raw text: Best quality Mikheev-style ending guessing rules, confidence level 90%.

Ending	Confidence	Class(es)	Frequency
ützer	0.918199	m4	10
urant	0.918199	n24	10
öhung	0.918199	f17	10
ptung	0.918199	f17	10

tgang	0.918199	m2	10
wort	0.91787	f17 n20 n22	246 5 14
bar	0.916662	m7 m8	65 4
üre	0.9161	f16	11
dit	0.9161	m1	11
gge	0.9161	f16	11
rve	0.9161	f16	11
end	0.915586	m1 f17 n20 n25	16 485 8 16
ve	0.915276	m7 f16	6 94
nche	0.914299	f16	10
omie	0.914299	f16	10
mide	0.914299	f16	10
lyse	0.914299	f16	10
buch	0.914299	n22	10
ativ	0.914299	n20	10
mant	0.914299	m8	10
rast	0.914299	m1	10
odil	0.914299	n20	10
rent	0.914299	m8	10
lenz	0.914299	f17	10
bube	0.914299	m7	10
mble	0.914299	n24	10
hein	0.914299	m1	10
urce	0.914299	f16	10
mium	0.914299	n28	10
omen	0.914299	n20	10
ferd	0.914299	n20	10
plar	0.914299	n20	10
werb	0.914299	m1	10
dat	0.912943	m8 n20	62 4
lucht	0.912129	f14 f17	4 58
u?e	0.911442	m7 f16	24 1
nke	0.910744	m7 f16	435 38
osse	0.909625	m7 f16	46 3
to	0.909465	m6 n24	1 26
mel	0.908807	m4 f16	245 21
rce	0.907837	f16	10
rch	0.907837	m2	10
yse	0.907837	f16	10
ord	0.907837	m1	10
erd	0.907837	n20	10
tom	0.907837	n20	10
äut	0.907837	n20	10
dil	0.907837	n20	10
äck	0.907837	n20	10
bund	0.906095	m1 m2	66 5
io	0.903747	n24	11
dy	0.903747	m6	11
if	0.903747	m1	11
ös	0.903747	m1	11
?e	0.903647	m7 f16	24 270
o?	0.902593	m1	1

		m2 n22	28 317

Table 36. Raw text: The last Mikheev-style ending guessing rules, confidence level 90%.

We currently use the 1789 ending guessing rules obtained through the raw text estimation. Other endings may be considered later. We would like to allow the generation of some ambiguous rules that predict up to *k* (e.g. *k*=5) different classes but do so with high confidence. We will be able to disambiguate later. While these rules are not categorical, it is always better to have an ambiguous guess than nothing. Remember that each stem has a set of acceptable morphological classes implicitly assigned during the previous stem refinement step. Consider a stem has an acceptable morphological class set {*m1*, *m9*, *f12*, *n20*, *n25*, *n28*} and matches an ambiguous rule that predicts the set {*m1*, *m7*, *m8*, *n20*, *n22*}. Then the real rule prediction is {*m1*, *n20*}, which is a reduction of the ambiguity set from 6 to 2 elements.

8.4.4 Cascade algorithm

```
0. Initialise the morphological class of all stems as UNKNOWN.
1. Consider the stems one-by-one and for each one do:
   1.1. If (is in the Stem Lexicon) ==> remove it; NEXT STEM;
   1.2. Try to split the word as a compound.
        If (split) => Assign the morphological class(es) of the last part; NEXT STEM;
   1.3. Try all endings predicting morphological classes, the longer one first:
        If (some ending matches) && (the prediction is compatible with the compatible class set)
            => Assign the morphological classes predicted; NEXT STEM;
2. END
```

8.4.5 Algorithm application

Table 37 shows the top unknown stems with the morphological information added. All the stems that cover at least three different word types are listed. The morphological information is of 4 different types:

KNOWN *stem*(*classes*) — the stem is already known

COMPOUND *stem*(*classes*) — at least one compound splitting has been found

ENDING RULE *ending*(*classes*) — an ending rule has been used

NO INFO — nothing of the above happened

Exactly one of these is listed. If more than one of these happened the highest label has been listed as it is considered to be more reliable. After the labels a list of all classes the rule is compatible with is listed in parentheses. In case of known stem, compound or ending rule the corresponding stem/ending is listed immediately after followed by the morphological class or classes it predicts. It is possible that there are more than one classes predicted by a single stem (see *Stadtteil*, Table 37) or more than one stems a compound can be split into (see *Gemeindehaushalt*, Table 37). In case of known stem it will be rejected at the subsequent step: no unknown word could have a known stem since all the words a known stem generates are known as well and are included in the Expanded Stem Lexicon. Table 37 contains one known stem: *Band* with morphological class *f15*. It will be rejected as possible stem candidate. We see that it is incompatible with the following acceptable morphological classes set. This incompatibility is very likely but not sure. On the other hand in case of compound or ending rule its prediction *must* be compatible with the following class set.

Unknown Stem	#	Words Covered by the Stem	Morphological Information
Ortsbeirat	5	{ ortsbeirat, ortsbeirates, ortsbeirats, ortsbeiräte, ortsbeiräten }	COMPOUND beirat(m2) rat(m2)(m2)
Gemeindehaushalt	**4**	**{ gemeindehaushalt, gemeindehaushalte, gemeindehaushaltes,**	**COMPOUND haushalt(m1) halt(m1)(m1 m2 m3 m3a**

		gemeindehaushalts }	**m9 n20 n21 n22 n25)**
Kinderarzt	4	{ kinderarzt, kinderarztes, kinderärzte, kinderärzten }	COMPOUND arzt(m2)(m2 n20a)
Kunstwerk	4	{ kunstwerk, kunstwerke, kunstwerken, kunstwerks }	COMPOUND werk(n20)(m1 m9 n20 n25)
Lebensjahr	4	{ lebensjahr, lebensjahren, lebensjahres, lebensjahrs }	COMPOUND jahr(n20)(m1 m9 n20 n25)
Ortsbezirk	4	{ ortsbezirk, ortsbezirke, ortsbezirken, ortsbezirks }	COMPOUND bezirk(m1)(m1 m9 n20 n25)
Stadtteil	**4**	**{ stadtteil, stadtteile, stadtteilen, stadtteils }**	**COMPOUND teil(m1,n20)(m1 m9 n20 n25)**
Wort	4	{ wort, worte, worten, wortes }	NO INFO (m1 m9 n20 n25)
Abend	3	{ abend, abende, abenden }	ENDING RULE abend(m1)(m1 m9 f12 n20 n25)
Ander	3	{ andere, anderen, anders }	NO INFO (m1 m9 n20 n25)
Anteilseigner	3	{ anteilseigner, anteilseignern, anteilseigners }	ENDING RULE ner(m4)(m4 m10 n23 n26 n30)
Arbeitsplatz	3	{ arbeitsplatz, arbeitsplätze, arbeitsplätzen }	COMPOUND platz(m2)(m2 f14 n20a)
Aufsichtsrat	3	{ aufsichtsrat, aufsichtsrates, aufsichtsrats }	COMPOUND rat(m2)(m1 m2 m3 m3a m9 n20 n21 n22 n25 n31)
Augenblick	3	{ augenblick, augenblicken, augenblicks }	COMPOUND blick(m1)(m1 m9 n20 n25)
Band	**3**	**{ bandes, bänder, bändern }**	**KNOWN band(f15)(m3 n22)**
Bau	3	{ bau, bauen, baus }	NO INFO (m1 m9 n20 n25)
Befreiungskampf	3	{ befreiungskampf, befreiungskampfes, befreiungskämpfer }	NO INFO (m3 n22)
Bensheim	3	{ bensheim, bensheimer, bensheims }	NO INFO (m3a n21)
Bernbach	3	{ bernbach, bernbacher, bernbachs }	NO INFO (m3a n21)
Biergarten	3	{ biergarten, biergartens, biergärten }	COMPOUND garten(m5)(m5 n23a)
Bildungsurlaub	3	{ bildungsurlaub, bildungsurlaube, bildungsurlauben }	NO INFO (m1 m9 f12 n20 n25)
Bildungsurlaube	3	{ bildungsurlaube, bildungsurlauben, bildungsurlauber }	NO INFO (m7)
Bo	3	{ bo, bose, boses }	NO INFO (m1a n27)
Brock	3	{ brock, brocks, bröcker }	NO INFO (m3 n22)
Bundesland	3	{ bundesland, bundesländer, bundesländern }	COMPOUND land(n22)(m3 n22)
Bürgerkrieg	3	{ bürgerkrieg, bürgerkrieges, bürgerkriegs }	COMPOUND krieg(m1)(m1 m2 m3 m3a m9 n20 n21 n22 n25 n31)
Edelstahlwerk	3	{ edelstahlwerke, edelstahlwerken, edelstahlwerkes }	COMPOUND werk(n20)(m1 m9 n20 n25)
Edelstahlwerke	3	{ edelstahlwerke, edelstahlwerken, edelstahlwerkes }	NO INFO (m4 m10 n23 n26 n30)
Eigentum	3	{ eigentum, eigentümer, eigentümern }	NO INFO (m3 n22)
Eigentümer	3	{ eigentümer, eigentümern, eigentümers }	NO INFO (m4 m10 n23 n26 n30)
Energieplan	3	{ energieplan, energieplaner, energieplans }	NO INFO (m3a n21)
Erfolgsrezept	3	{ erfolgsrezept, erfolgsrezepten, erfolgsrezepts }	COMPOUND rezept(n20)(m1 m9 n20 n25)
Flörsheim	3	{ flörsheim, flörsheimer, flörsheims }	NO INFO (m3a n21)
Geist	3	{ geist, geiste, geistes }	NO INFO (m1 m2 m3 m3a m9 n20 n20a n21 n22 n25)
Georg	3	{ georg, george, georges }	NO INFO (m1 m2 m3 m3a m9 n20 n20a n21 n22 n25)
Geschehen	3	{ geschehen, geschehene, geschehens }	NO INFO (m1 m2 m3 m3a m9 n20 n21 n22 n25)
Grundrecht	3	{ grundrecht, grundrechte,	COMPOUND recht(n20)(m1

		grundrechts }	m2 m3 m3a m9 n20 n21 n22 n25)
Gruppenspiel	3	{ gruppenspiel, gruppenspiele, gruppenspielen }	COMPOUND spiel(n20)(m1 m9 f12 n20 n25)
Hanau	3	{ hanau, hanauer, hanaus }	NO INFO (m3a n21)
Herman	3	{ herman, hermann, hermanns }	NO INFO (m7a)
Hochmoor	3	{ hochmoor, hochmoore, hochmooren }	COMPOUND moor(n20)(m1 m9 f12 n20 n25)
Hunderte	3	{ hunderte, hunderten, hunderter }	NO INFO (m7)
Idyll	3	{ idylle, idyllen, idylls }	NO INFO (m1 m9 n20 n25)
Ing	3	{ ing, inge, inger }	NO INFO (m3a n21)
Jugendzentr	3	{ jugendzentren, jugendzentrum, jugendzentrums }	COMPOUND zentr(n28)(n28)
Karnevalverein	3	{ karnevalverein, karnevalvereine, karnevalvereinen }	COMPOUND verein(m1)(m1 m9 f12 n20 n25)
Kinderarzte	3	{ kinderarztes, kinderärzte, kinderärzten }	NO INFO (m5 n23a)
Kindergarten	3	{ kindergarten, kindergartens, kindergärten }	COMPOUND garten(m5)(m5 n23a)
Krankenhaus	3	{ krankenhaus, krankenhäuser, krankenhäusern }	COMPOUND haus(n22)(m3 n22)
Kreisvorsitzende	3	{ kreisvorsitzende, kreisvorsitzenden, kreisvorsitzender }	COMPOUND vorsitzende(m7,f16)(m7)
Langenhain	3	{ langenhain, langenhainer, langenhains }	NO INFO (m3a n21)
Lebenslauf	3	{ lebenslauf, lebenslaufes, lebensläufe }	COMPOUND lauf(m2)(m2 n20a)
Leut	3	{ leut, leute, leuten }	NO INFO (m1 m9 f12 n20 n25)
Munch	3	{ munch, munchs, münchen }	NO INFO (m2)
Musikzug	3	{ musikzug, musikzugs, musikzüge }	COMPOUND zug(m2)(m2)
Mörlenbach	3	{ mörlenbach, mörlenbachern, mörlenbachs }	NO INFO (m3a n21)
Name	3	{ name, namen, namens }	NO INFO (m7a)
Ortsbeirate	3	{ ortsbeirates, ortsbeiräte, ortsbeiräten }	NO INFO (m5 n23a)
Ost	3	{ ost, oster, ostern }	NO INFO (m3a n21)
Papp	3	{ papp, pappe, pappen }	NO INFO (m1 m9 f12 n20 n25)
Programmheft	3	{ programmheft, programmhefte, programmheften }	COMPOUND heft(n20)(m1 m9 f12 n20 n25)
Punkt	3	{ punkt, punkte, punkten }	ENDING RULE punkt(m1)(m1 m9 f12 n20 n25)
Regenwald	3	{ regenwald, regenwaldes, regenwälder }	COMPOUND wald(m3)(m3 n22)
Schmitt	3	{ schmitt, schmitten, schmitts }	ENDING RULE itt(m1)(m1 m9 n20 n25)
Schuldenberg	3	{ schuldenberge, schuldenberges, schuldenbergs }	COMPOUND berg(m1)(m1 m2 m3 m3a m9 n20 n21 n22 n25)
Sitzplatz	3	{ sitzplatz, sitzplätze, sitzplätzen }	COMPOUND platz(m2)(m2 f14 n20a)
Spd-fraktionsvorsitzende	3	{ spd-fraktionsvorsitzende, spd-fraktionsvorsitzenden, spd-fraktionsvorsitzender }	NO INFO (m7)
Spielplatz	3	{ spielplatz, spielplätze, spielplätzen }	COMPOUND platz(m2)(m2 f14 n20a)
Spieltag	3	{ spieltag, spieltage, spieltagen }	COMPOUND tag(m1)(m1 m9 f12 n20 n25)
Sportplatz	3	{ sportplatz, sportplätze, sportplätzen }	COMPOUND platz(m2)(m2 f14 n20a)
Sportverein	3	{ sportverein, sportvereine, sportvereins }	COMPOUND verein(m1)(m1 m2 m3 m3a m9 n20 n21 n22 n25)
Stadtteilparlament	3	{ stadtteilparlament, stadtteilparlamentes, stadtteilparlaments }	COMPOUND parlament(n20)(m1 m2 m3 m3a m9 n20 n21 n22

			n25 n31)
Stadtverordnete	3	{ stadtverordnete, stadtverordneten, stadtverordneter }	COMPOUND verordnete(m7)(m7)
Stahlwerk	3	{ stahlwerk, stahlwerke, stahlwerker }	NO INFO (m3a n21)
Stra?enbauamt	3	{ stra?enbauamt, stra?enbauamtes, stra?enbauamts }	COMPOUND amt(n22)(m1 m2 m3 m3a m9 n20 n21 n22 n25 n31)
Tagebuch	3	{ tagebuch, tagebuchs, tagebüchern }	COMPOUND buch(n22)(m3 n22)
Tarifvertrag	3	{ tarifvertrag, tarifvertrags, tarifverträgen }	COMPOUND vertrag(m2)(m2)
Tibet	3	{ tibet, tibeter, tibetern }	NO INFO (m3a n21)
Tod	3	{ tod, tode, todes }	NO INFO (m1 m2 m3 m3a m9 n20 n20a n21 n22 n25)
Vereinsheim	3	{ vereinsheim, vereinsheimen, vereinsheims }	COMPOUND heim(n20)(m1 m9 n20 n25)
Verwaltungshaushalt	3	{ verwaltungshaushalt, verwaltungshaushaltes, verwaltungshaushalts }	COMPOUND haushalt(m1)(m1 m2 m3 m3a m9 n20 n21 n22 n25 n31)
West	3	{ west, weste, westen }	NO INFO (m1 m9 f12 n20 n25)
Worte	3	{ worte, worten, wortes }	NO INFO (m4 m10 n23 n26 n30)
Zehntausend	3	{ zehntausend, zehntausende, zehntausenden }	NO INFO (m1 m9 f12 n20 n25)

Table 37. Unknown stems with morphological information. (NEGRA corpus)

8.5 Word types clusterisation (stem coverage)

For each hypothetical stem we keep information which word types it is supposed to cover. After the stem refinements step we are sure that each stem is compatible with the word types it is supposed to cover and that there exists at least one morphological class that could generate them all given the stem. During the next step we obtained some additional information regarding the stems as a result of morphological analysis.

We thus obtained a complex structure, which we can think of as a bi-partition graph where the vertices are either stems or word types and the edges link each stem to the word type that it is supposed to cover. It is clear that in the general case this is a multigraph since each stem could be generated by more than one word and each word may be covered by several different stems. Our goal is to select some of the stems making the stem coverage of the word types. We try to select some of the stems in a way that:

1) Each word is covered by exactly one stem. (pigeon hole principle)
2) The stem covers as much word types as possible
3) The covered word types set being equal, a stem with more reliable morphological information attached is preferred. This means we prefer a stem that could be classified using an ending guessing rule to one without any morphological information and a stem that has been recognised as a compound to stem that is covered by ending guessing rule. (The known stems are simply rejected, see above).
4) All other being equal, a longer stem is preferred.

The first consideration is a simplification. In fact it is possible that two different stems share the same word form but this is quite unlikely and for the moment we prefer to simply reject this possibility in order to keep the model simpler. The word will still be attached to a stem but to exactly one among the possible ones. We would like to stress that this simplification assumption is different from the one that permits us to reject the known stems as candidates for unknown words. In the latter case this was motivated by the fact that all the word forms a known stem could generate are present in the Expanded Stem Lexicon. In the present case this is not a simplification of the same type but rather exploitation of a known property.

The second criterion is based on observation that if there is a stem and corresponding morphological class that could cover certain set of word forms then it is most likely that this is the correct stem and not a candidate that covers just a subset of these. Though, we did not performed formal tests and just observed some random samples, which means this criterion have to be justified further in real experiments.

The third criterion is clearer: if we have to choose between two stems covering the same set of word forms and one of these is recognised as a compound whose last part is a known stem than it is more likely that this is the correct stem and not the one that we have less information about. That is because to recognise a word as a compound is very restrictive (we want to find all the words it is composed of in the corresponding lexicons and with the appropriate parts of speech, see above). The incorrect identification of a compound is quite unlikely since we perform this checking while on the other hand the compounding process is very common and very powerful. The same way we prefer a word having a known ending that predicts some morphological class according to a rule to one with unknown ending. The motivation behind is that the known endings that enter in ending-guessing rules are among the most frequent ones. The more frequent an ending the more likely it is to be the correct one if more than one possibility is present. Anyway, while these considerations may sound somewhat intuitive they have to be tested formally.

The fourth rule just tries to keep the things more conservative. Remember that the stem by definition was the longest common prefix shared by all forms of the same word. (In fact there were some particularities with the umlauts, see above). Given all the word forms the stem identification is straightforward. But if we have just part of them it may be impossible since more than one morphological class may cover a set of 2 or 3 word forms. But since the stem is supposed to be the longest common prefix of all word forms we prefer to be as near to this definition as possible even in case of missing word forms and limited information available (in the general case we do not know for sure neither the case nor the number of a particular word form). We thus prefer the longest stem among the candidates, all other properties being equal.

How we solve the coverage problem? We do this in two steps: sorting and selection. We sort the stem candidates the better candidates first and then we perform an additional pass through the sorted stem list during which we either select or reject each of the stems. The stems are sorted by three criteria:

1) word types covered count
2) morphological information available
3) stem length

The most important criterion is the word types covered count. The more word types a stem covers the better candidate it is. In case two stems cover the same count of word types (but not necessarily exactly the same word type set) we look at the morphological information available. We prefer the stems that have been decomposed as compounds to those whose endings are known and predict a morphological class through ending guessing rules, and those with known endings to those without any morphological information available. In case two stems cover the same word types count and have the same morphological information (both are compounds, both have known endings or neither applies to both at the same time) then we look at the stem length and put the longer stem before the shorter one. We then go through the stems and either select or reject it. We accept a stem only if all the word types it covers have not been assigned to another stem till now and reject it otherwise. We thus do not allow a word to be covered by more than one stem.

Algorithm

1. Initialise each word type as uncovered.

2. Sort the stems by word types that generated the stem count (in decreasing order), then by morphological information available (compound, ending rule, nothing) and then by stem length (decreasing order).

3. Consider the sorted stems one-by-one:

If at least one of the corresponding word types has been covered — reject the stem, otherwise — accept it and mark all word types it covers as covered.

4. End.

Table 38 contains the top selected unknown stems together with the corresponding morphological information available from the previous step. Compare it to Table 37 that contains the stem list before the selection. The stem *Wort* has been accepted and the candidate *Worte* has been rejected since the first one covers 4 word types while the second does just 3. The stem *Edelstahlwerk* has been accepted while *Edelstahlwerke* has been rejected. They cover exactly the same word type set and thus sets with the same members count which means they are equal according to the first criterion. But the shorter stem candidate has been recognised as a compound while the second one did not and this decided the choice.

Selected Stem	#	Words Covered by the Stem	Morphological Information
Ortsbeirat	5	{ Ortsbeirat, Ortsbeirates, Ortsbeirats, Ortsbeiräte, Ortsbeiräten }	COMPOUND beirat(m2) rat(m2)(m2)
Gemeindehaushalt	4	{ Gemeindehaushalt, Gemeindehaushalte, Gemeindehaushaltes, Gemeindehaushalts }	COMPOUND haushalt(m1) halt(m1)(m1 m2 m3 m3a m9 n20 n21 n22 n25)
Kinderarzt	4	{ Kinderarzt, Kinderarztes, Kinderärzte, Kinderärzten }	COMPOUND arzt(m2)(m2 n20a)
Kunstwerk	4	{ Kunstwerk, Kunstwerke, Kunstwerken, Kunstwerks }	COMPOUND werk(n20)(m1 m9 n20 n25)
Lebensjahr	4	{ Lebensjahr, Lebensjahren, Lebensjahres, Lebensjahrs }	COMPOUND jahr(n20)(m1 m9 n20 n25)
Ortsbezirk	4	{ Ortsbezirk, Ortsbezirke, Ortsbezirken, Ortsbezirks }	COMPOUND bezirk(m1)(m1 m9 n20 n25)
Stadtteil	4	{ Stadtteil, Stadtteile, Stadtteilen, Stadtteils }	COMPOUND teil(m1,n20)(m1 m9 n20 n25)
Wort	4	{ Wort, Worte, Worten, Wortes }	NO INFO (m1 m9 n20 n25)
Abend	3	{ Abend, Abende, Abenden }	ENDING RULE abend(m1)(m1 m9 f12 n20 n25)
Ander	3	{ Andere, Anderen, Anders }	NO INFO (m1 m9 n20 n25)
Anteilseigner	3	{ Anteilseigner, Anteilseignern, Anteilseigners }	ENDING RULE ner(m4)(m4 m10 n23 n26 n30)
Arbeitsplatz	3	{ Arbeitsplatz, Arbeitsplätze, Arbeitsplätzen }	COMPOUND platz(m2)(m2 f14 n20a)
Aufsichtsrat	3	{ Aufsichtsrat, Aufsichtsrates, Aufsichtsrats }	COMPOUND rat(m2)(m1 m2 m3 m3a m9 n20 n21 n22 n25 n31)
Augenblick	3	{ Augenblick, Augenblicken, Augenblicks }	COMPOUND blick(m1)(m1 m9 n20 n25)
Bau	3	{ Bau, Bauen, Baus }	NO INFO (m1 m9 n20 n25)
Befreiungskampf	3	{ Befreiungskampf, Befreiungskampfes, Befreiungskämpfer }	NO INFO (m3 n22)
Bensheim	3	{ Bensheim, Bensheimer, Bensheims }	NO INFO (m3a n21)
Bernbach	3	{ Bernbach, Bernbacher, Bernbachs }	NO INFO (m3a n21)
Biergarten	3	{ Biergarten, Biergartens, Biergärten }	COMPOUND garten(m5)(m5 n23a)
Bildungsurlaube	3	{ Bildungsurlaube, Bildungsurlauben, Bildungsurlauber }	NO INFO (m7)
Bo	3	{ Bo, Bose, Boses }	NO INFO (m1a n27)
Brock	3	{ Brock, Brocks, Bröcker }	NO INFO (m3 n22)
Bundesland	3	{ Bundesland, Bundesländer, Bundesländern }	COMPOUND land(n22)(m3 n22)
Bürgerkrieg	3	{ Bürgerkrieg, Bürgerkrieges, Bürgerkriegs }	COMPOUND krieg(m1)(m1 m2 m3 m3a m9 n20 n21 n22 n25 n31)
Edelstahlwerk	3	{ Edelstahlwerke, Edelstahlwerken, Edelstahlwerkes }	COMPOUND werk(n20)(m1 m9 n20 n25)
Eigentümer	3	{ Eigentümer, Eigentümern, Eigentümers }	NO INFO (m4 m10 n23 n26 n30)
Energieplan	3	{ Energieplan, Energieplaner, Energieplans }	NO INFO (m3a n21)
Erfolgsrezept	3	{ Erfolgsrezept, Erfolgsrezepten,	COMPOUND rezept(n20)(m1

		Erfolgsrezepts }	m9 n20 n25)
Flörsheim	3	{ Flörsheim, Flörsheimer, Flörsheims }	NO INFO (m3a n21)
Geist	3	{ Geist, Geiste, Geistes }	NO INFO (m1 m2 m3 m3a m9 n20 n20a n21 n22 n25)
Georg	3	{ Georg, George, Georges }	NO INFO (m1 m2 m3 m3a m9 n20 n20a n21 n22 n25)
Geschehen	3	{ Geschehen, Geschehene, Geschehens }	NO INFO (m1 m2 m3 m3a m9 n20 n21 n22 n25)
Grundrecht	3	{ Grundrecht, Grundrechte, Grundrechts }	COMPOUND recht(n20)(m1 m2 m3 m3a m9 n20 n21 n22 n25)
Gruppenspiel	3	{ Gruppenspiel, Gruppenspiele, Gruppenspielen }	COMPOUND spiel(n20)(m1 m9 f12 n20 n25)
Hanau	3	{ Hanau, Hanauer, Hanaus }	NO INFO (m3a n21)
Herman	3	{ Herman, Hermann, Hermanns }	NO INFO (m7a)
Hochmoor	3	{ Hochmoor, Hochmoore, Hochmooren }	COMPOUND moor(n20)(m1 m9 f12 n20 n25)
Hunderte	3	{ Hunderte, Hunderten, Hunderter }	NO INFO (m7)
Idyll	3	{ Idylle, Idyllen, Idylls }	NO INFO (m1 m9 n20 n25)
Ing	3	{ Ing, Inge, Inger }	NO INFO (m3a n21)
Jugendzentr	3	{ Jugendzentren, Jugendzentrum, Jugendzentrums }	COMPOUND zentr(n28)(n28)
Karnevalverein	3	{ Karnevalverein, Karnevalvereine, Karnevalvereinen }	COMPOUND verein(m1)(m1 m9 f12 n20 n25)
Kindergarten	3	{ Kindergarten, Kindergartens, Kindergärten }	COMPOUND garten(m5)(m5 n23a)
Krankenhaus	3	{ Krankenhaus, Krankenhäuser, Krankenhäusern }	COMPOUND haus(n22)(m3 n22)
Kreisvorsitzende	3	{ Kreisvorsitzende, Kreisvorsitzenden, Kreisvorsitzender }	COMPOUND vorsitzende(m7,f16)(m7)
Langenhain	3	{ Langenhain, Langenhainer, Langenhains }	NO INFO (m3a n21)
Lebenslauf	3	{ Lebenslauf, Lebenslaufes, Lebensläufe }	COMPOUND lauf(m2)(m2 n20a)
Leut	3	{ Leut, Leute, Leuten }	NO INFO (m1 m9 f12 n20 n25)
Munch	3	{ Munch, Munchs, München }	NO INFO (m2)
Musikzug	3	{ Musikzug, Musikzugs, Musikzüge }	COMPOUND zug(m2)(m2)
Mörlenbach	3	{ Mörlenbach, Mörlenbachern, Mörlenbachs }	NO INFO (m3a n21)
Name	3	{ Name, Namen, Namens }	NO INFO (m7a)
Ost	3	{ Ost, Oster, Ostern }	NO INFO (m3a n21)
Papp	3	{ Papp, Pappe, Pappen }	NO INFO (m1 m9 f12 n20 n25)
Programmheft	3	{ Programmheft, Programmhefte, Programmheften }	COMPOUND heft(n20)(m1 m9 f12 n20 n25)
Punkt	3	{ Punkt, Punkte, Punkten }	ENDING RULE punkt(m1)(m1 m9 f12 n20 n25)
Regenwald	3	{ Regenwald, Regenwaldes, Regenwälder }	COMPOUND wald(m3)(m3 n22)
Schmitt	3	{ Schmitt, Schmitten, Schmitts }	ENDING RULE itt(m1)(m1 m9 n20 n25)
Schuldenberg	3	{ Schuldenberge, Schuldenberges, Schuldenbergs }	COMPOUND berg(m1)(m1 m2 m3 m3a m9 n20 n21 n22 n25)
Sitzplatz	3	{ Sitzplatz, Sitzplätze, Sitzplätzen }	COMPOUND platz(m2)(m2 f14 n20a)
Spd-fraktionsvorsitzende	3	{ Spd-Fraktionsvorsitzende, Spd-Fraktionsvorsitzenden, Spd-Fraktionsvorsitzender }	NO INFO (m7)
Spielplatz	3	{ Spielplatz, Spielplätze, Spielplätzen }	COMPOUND platz(m2)(m2 f14 n20a)
Spieltag	3	{ Spieltag, Spieltage, Spieltagen }	COMPOUND tag(m1)(m1 m9 f12 n20 n25)
Sportplatz	3	{ Sportplatz, Sportplätze, Sportplätzen }	COMPOUND platz(m2)(m2 f14 n20a)
Sportverein	3	{ Sportverein, Sportvereine, Sportvereins }	COMPOUND verein(m1)(m1 m2 m3 m3a m9 n20 n21 n22 n25)

Stadtteilparlament	3	{ Stadtteilparlament, Stadtteilparlamentes, Stadtteilparlaments }	COMPOUND parlament(n20)(m1 m2 m3 m3a m9 n20 n21 n22 n25 n31)
Stadtverordnete	3	{ Stadtverordnete, Stadtverordneten, Stadtverordneter }	COMPOUND verordnete(m7)(m7)
Stahlwerk	3	{ Stahlwerk, Stahlwerke, Stahlwerker }	NO INFO (m3a n21)
Stra?enbauamt	3	{ Stra?enbauamt, Stra?enbauamtes, Stra?enbauamts }	COMPOUND amt(n22)(m1 m2 m3 m3a m9 n20 n21 n22 n25 n31)
Tagebuch	3	{ Tagebuch, Tagebuchs, Tagebüchern }	COMPOUND buch(n22)(m3 n22)
Tarifvertrag	3	{ Tarifvertrag, Tarifvertrags, Tarifverträgen }	COMPOUND vertrag(m2)(m2)
Tibet	3	{ Tibet, Tibeter, Tibetern }	NO INFO (m3a n21)
Tod	3	{ Tod, Tode, Todes }	NO INFO (m1 m2 m3 m3a m9 n20 n20a n21 n22 n25)
Vereinsheim	3	{ Vereinsheim, Vereinsheimen, Vereinsheims }	COMPOUND heim(n20)(m1 m9 n20 n25)
Verwaltungshaushalt	3	{ Verwaltungshaushalt, Verwaltungshaushaltes, Verwaltungshaushalts }	COMPOUND haushalt(m1)(m1 m2 m3 m3a m9 n20 n21 n22 n25 n31)
West	3	{ West, Weste, Westen }	NO INFO (m1 m9 f12 n20 n25)
Zehntausend	3	{ Zehntausend, Zehntausende, Zehntausenden }	NO INFO (m1 m9 f12 n20 n25)

Table 38. ***Selected*** **unknown stems** together with the morphological information available till now. (NEGRA corpus)

8.6 Deterministic context exploitation

Historically, the System has been implemented as a set of different modules each of which has been tested separately and just then linked to some of the others. All the steps described above are linked together as parts of the current version of the System. But there are still some separate modules that although have been developed and tested already are not yet linked. Since these are very important modules that will undoubtedly be added to the System we will describe them below. In fact the probabilistic context exploitation module, which is based on word type context vectors, was the very first module we implemented. But it has been left out anyway since it is the last step to be performed by the System. The deterministic context information exploitation module is to be linked between the stem refinement and the morphological analysis steps but has been left out as well since it is very similar to the probabilistic one and we decided it would be much easier to link them at the same time at a later stage.

The context information is exploited in both deterministic and probabilistic way. These could be applied at the same time but it is better if this is done separately as described above. The purpose of the deterministic context exploitation is to check whether a particular morphological class assigned to a stem is acceptable looking at the contexts of the word types it is supposed to cover. The idea is that some very frequent closed class words are highly predictive in what about the case and/or gender and/or number of the word token they precede. For example the articles in German are put before the noun they modify and change by both number and case, see Table 39. Th? article *das* predicts the following noun is neuter/singular/nominative or neuter/singular/accusative, while *den* predicts masculine/singular/accusative or plural/dative for all genders. Unlike other languages (e.g. French) German has *no* separate plural forms for the different genders.

Consider we have a stem candidate, set of word types it is supposed to cover and a set of acceptable morphological classes obtained during the stem refinement step. We would like to check whether each of the morphological classes is acceptable looking at the context. We check the classes one-by-one. Once we have chosen a class to check it automatically fixes the possible stem gender and from there — the gender of all word types it is supposed to cover. This implies as well some constraints on both the number and case for each word type. As we saw above each definite article form (the same applies to other kinds of predictors) implies its own constraints on the subsequent word token. What we have to do is to check whether the context constraints due to a particular word token match the constraints for the corresponding word type.

Case	Singular			Plural
	Masculine	Feminine	Neuter	
Nominative	der	die	das	die
Genitive	des	der	des	der
Dative	dem	der	dem	den
Accusative	den	die	das	die

Table 39. German definite article declination.

Let us take as an example the stem *Ost*, which is supposed to cover the word type set { *Ost*, *Oster*, *Ostern* }. There are two possible morphological classes: *m3a* and *n21*. Consider we investigate the possibility that the morphological class *m3a* is acceptable. We investigate the word types one-by-one and for each of them look at the contexts of all its corresponding word tokens. Suppose we see the article *der* before a particular word token of *Ost*. This is a zero-ending word type form of the stem *Ost*. Looking at the inflections of the morphological class *m3a* we can conclude this is nominative/singular, dative/singular or accusative/singular. Looking at the predictor *der* we see it could be nominative/singular/masculine, genitive/singular/feminine, dative/singular/feminine and genitive/plural for all genders. We check the intersection of the two sets:

{ nom/sg/mas, dat/sg/mas, akk/sg/mas }
{ nom/sg/mas, gen/pl/mas, gen/pl/fem, gen/pl/neu }

and find it is non-empty: { nom/sg/mas }. This means *der Ost* is explained by the morphological class *m3a* as nominative/singular/masculine and we cannot reject *m3a* as candidate. If there were other predictors for this or for other word type among the ones the stem is supposed to cover we would check them as well and conclude *m3a* is acceptable only if all they can be explained by the morphological class.

Looking at the morphological class *n21* for the same combination *der Ost* we obtain the sets:

{ nom/sg/mas, dat/sg/mas, akk/sg/mas }
{ gen/pl/mas, gen/pl/fem, gen/pl/neu }

This time they are incompatible: the first set contains only singular forms while the second one contains only plural forms. This means the combination *der Ost* cannot be explained by the morphological class *n21* and it has to be rejected.

We will not enter in more details here and will leave them for the following section where we explain the probabilistic vectors creation.

Remark

In fact much richer and much more reliable context information could be available through a POS tagger. This is the approach adopted by several similar systems like Morphy but we are not willing to do so at this moment. The application of a POS tagger to unknown words is unreliable and could introduce an uncontrollable amount of errors. That is why we prefer to lose some potentially useful cues but to be sure that the ones we use are quite reliable. From our point of view it is better to output a ranked list of several possible morphological classes that will contain the correct one than to lose it as a result of an incorrect disambiguation. Note that our morphological analyser is designed to be used as a part of a bigger system rather than as a stand-alone application. This system could refine its output through a POS tagger or use it to leverage the results obtained from a POS tagger. ?

8.7 Word types context vector creation

The probabilistic context exploitation is based on word type vectors. Below we explain the idea in more details and give several example vectors of this type.

After the word types have been covered by stems we build vectors for each separate word type. The vector has 24 (3×2×4) coordinates and can be thought of as a three-dimensional cube measured by: gender (3), number (2) and case (4). After the vector creation phase each word type whose stem has not been classified in a deterministic way during phase 3 will obtain its own vector. Note that we create vectors for the word types and *not* for the stems. In the general case the vector coordinates will sum to one and can be thought of as probabilities and the vectors — as probability distributions. In case no predictors were present in the text for a specific word type it will not have vector (all vector coordinates will be 0).

How are the vectors created? First the vectors for each word are initialised with zeroes. Then a pass through the text is performed and contextual predictor information is collected. Each time we encounter a predictor its vector is used to modify the word type vector corresponding to the following word token. Several types of predictors could be used:

- articles:
  - ✓ das, dem, den, der, des, die;
  - ✓ ein, eine, einem, einen, einer, eines;
  - ✓ kein, keine, keinem, keinen, keiner, keines.
- prepositions (see Table 40)
- pronouns: possessive, demonstrative, indefinite (not used currently)

The articles are some of the most important predictors because they are among the most frequent words and are very likely to be used before an unknown word. They can provide useful information regarding gender (G), number (N) and case (C) of a specific word token. We use this information to build a vector showing how likely is that the word type this token belongs to take each G/N/C combination. Since the same article form can designate different G/N/C combinations and they are not equally likely we estimated them from the NEGRA corpus. Figure 32 shows the frequency distribution for the German articles. It is important to say that the NEGRA corpus is only partially annotated with morphological information. Sometimes some of the information is missing. Only the first 60,000 words out of 170,000 have been wholly annotated. The rest are either not annotated at all or annotated only partially and at least one of the morphological characteristics is missing: case, gender or number. We tried to use this incomplete information distributing the occurrence frequencies among all the G/N/C combinations they cover. For example if we see *der* annotated as *nom/sg* we have to distribute the occurrence among all the three genders. If we see it as only *nom* we have 6 possibilities. Clearly, this introduced a lot of noise and we finally decided to use the complete annotations only.

```
das         440
            NOM     GEN     DAT     AKK            NOM     GEN     DAT     AKK
     M      0       0       0       0              0       0       0       0
     F      0       0       0       0              0       0       0       0
     N      245     0       0       195            0       0       0       0
dem         392
            NOM     GEN     DAT     AKK            NOM     GEN     DAT     AKK
     M      0       0       250     0              0       0       0       0
     F      0       0       0       0              0       0       0       0
     N      0       0       142     0              0       0       0       0
den         318
            NOM     GEN     DAT     AKK            NOM     GEN     DAT     AKK
     M      0       0       0       318            0       0       0       0
     F      0       0       0       0              0       0       0       0
     N      0       0       0       0              0       0       0       0
der         1433
            NOM     GEN     DAT     AKK            NOM     GEN     DAT     AKK
     M      427     0       0       0              0       0       0       0
     F      0       405     601     0              0       0       0       0
```

	N	0	0	0	0	0	0	0	0
des		337							
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	0	211	0	0	0	0	0	0
	F	0	0	0	0	0	0	0	0
	N	0	126	0	0	0	0	0	0
die		910							
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	0	0	0	0	0	0	0	0
	F	456	0	0	454	0	0	0	0
	N	0	0	0	0	0	0	0	0
ein		294							
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	108	0	0	0	0	0	0	0
	F	0	0	0	0	0	0	0	0
	N	63	0	3	120	0	0	0	0
eine		271							
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	0	0	0	0	0	0	0	0
	F	107	0	0	164	0	0	0	0
	N	0	0	0	0	0	0	0	0
einem		122							
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	0	0	61	0	0	0	0	0
	F	0	0	0	0	0	0	0	0
	N	0	0	61	0	0	0	0	0
einen		118							
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	0	0	0	116	0	0	0	0
	F	0	0	2	0	0	0	0	0
	N	0	0	0	0	0	0	0	0
einer		177							
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	15	0	0	0	0	0	0	0
	F	0	46	116	0	0	0	0	0
	N	0	0	0	0	0	0	0	0
eines		65							
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	0	35	0	2	0	0	0	0
	F	0	0	0	0	0	0	0	0
	N	4	22	0	2	0	0	0	0
kein		22							
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	8	0	0	0	0	0	0	0
	F	0	0	0	0	0	0	0	0
	N	7	0	0	7	0	0	0	0
keine		28							
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	0	0	0	0	0	0	0	0
	F	13	0	0	15	0	0	0	0
	N	0	0	0	0	0	0	0	0
keinem			1						
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	0	0	0	0	0	0	0	0
	F	0	0	0	0	0	0	0	0
	N	0	0	1	0	0	0	0	0
keinen			13						
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	0	0	0	13	0	0	0	0
	F	0	0	0	0	0	0	0	0
	N	0	0	0	0	0	0	0	0
keiner			6						
		NOM	GEN	DAT	AKK	NOM	GEN	DAT	AKK
	M	4	0	0	0	0	0	0	0
	F	0	0	2	0	0	0	0	0
	N	0	0	0	0	0	0	0	0

Figure 32. Frequency distribution for the German articles from the NEGRA corpus.

By normalising we obtain the vectors corresponding to each of the articles (in fact some of these can be used as other parts of speech e.g. pronouns but we do not distinguish between these). In fact these are true maximum likelihood estimates of the probability distribution P(G/N/C | context).

What is important here is that even for the determiners, which are supposed to be very frequent, we failed to get reliable maximum likelihood estimations. In plus some of the determiners were not met (or were met but had no morphological tags) at all in the NEGRA corpus and we thus were unable to build any vectors for them: *keines*. We built artificial G/N/C vectors for them taking the possible G/N/C combinations from a German grammar and assuming they are equally likely.

The prepositions are important since they can provide information about the case. Since the prepositions are not so frequent and we want to be sure we have reliable predictions, the vectors for the prepositions were build according to the German grammar. There are three types of predictors: for Genitive, for Dative, for Accusative and a fourth group for both Dative and Accusative (see Table 40).

Case	Prepositions
Genitive	abseits, abzüglich, anfangs, angesichts, anhand, anläßlich, anstatt, anstelle, antwortlich, aufgrund, ausgangs, ausschließlich, außerhalb, behufs, beiderseits, betreffs, bezüglich, diesseits, eingangs, einschließlich, exklusive, halber, hinsichtlich, infolge, inklusive, inmitten, innerhalb, jenseits, kraft, längs, längsseits, laut, mangels, mittels, namens, oberhalb, rücksichtlich, seitens, seitlich, seitwärts, statt, trotz, um-willen, unbeschadet, unerachtet, unfern, ungeachtet, unterhalb, unweit, vermittels, vermöge, vorbehaltlich, während, wegen, vonwegen, zeit, zufolge, zugunsten, zuliebe, zuungunsten, zuzüglich, zwecks
Dative	ab, aus, außer, bei, binnen, dank, entgegen, fern, gegenüber, gemäß, mit, mitsamt, nach, nächst, nahe, nebst, ob, samt, seit, von, zu, zunächst, zuwider
Accusative	bis, durch, entlang, für, gegen, gen, ohne, per, sonder, um, wider
Dative or Accusative	an, auf, hinter, in, neben, über, unter, vor, zwischen

Table 40. German prepositions and the case(s) they predict.

We assumed that the only thing that matters regarding the prepositions above is the case and we assumed uniform distributions across the gender and number given the case. Thus, we built artificial statistics (again by normalising we get a probability distribution or a G/N/C vector). Figure 33 shows some examples of this artificial frequency distribution (without normalisation).

```
abseits              6
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M       0       1       0       0               0       1       0       0
     F       0       1       0       0               0       1       0       0
     N       0       1       0       0               0       1       0       0
ob           6
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M       0       0       1       0               0       0       1       0
     F       0       0       1       0               0       0       1       0
     N       0       0       1       0               0       0       1       0
für          6
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M       0       0       0       1               0       0       0       1
     F       0       0       0       1               0       0       0       1
     N       0       0       0       1               0       0       0       1
in           12
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M       0       0       1       1               0       0       1       1
     F       0       0       1       1               0       0       1       1
     N       0       0       1       1               0       0       1       1
```

Figure 33. Artificial frequency distributions for the German prepositions.

The next important group is formed by the pronouns. For the pronouns we estimated the distribution and selected the best ones. The information from this source is unreliable due to insufficient statistics and to the fact they can be used not only as noun modifiers but also instead of nouns and thus are excluded from the baseline experiments. But they will be considered in the subsequent experiments.

How the predictor information is used and how the word types vectors are created? We go through the corpus and if we encounter a predictor we remember it. In case we encounter an acceptable noun after it, we update its vector with information contained in the predictor. The acceptable noun is the first noun following the predictor, not necessarily immediately. The predictor is discarded in case a sentence end `.!?` or boundary mark `(),.:;-/<>` is encountered. In case we encounter another predictor of the same type before encountering a noun, we discard the old predictor and remember the new one.

The update procedure consists in adding the weighted G/N/C vector of the predictor to the corresponding word type vector. The weighting is reversibly proportional to the non-zero predictor vector coordinates: the more non-zero values, the lower the quality of the prediction and thus the lower the weight. For the baseline model we scale the predictor vector by 1/NZ, where NZ is the non-zero vector coordinates count.

Since the prepositions are likely to introduce a high noise level because of the shortage of constraints on both number and gender, we did not use them as predictors taken alone but only in combination with an article. Only the combinations of preposition followed by an article or article alone are considered rejecting the others. In case the prepositions and the article have incompatible distributions (no G/N/C vector coordinate exists for which they are both non-zero) the preposition is ignored and only the article G/N/C vector is taken into account. In case the vectors of the preposition-article couple are compatible (their dot product is non-zero) they are multiplied coordinate by coordinate and the vector thus obtained is normalised to get pseudo probabilities as coordinates. The vector is then weighted by a constant factor (e.g. 3 or 5) and no other scaling (e.g. dependent on the non-zero elements) is applied.

There is something in plus we skipped above to make the things simpler to explain. Remember that when we analysed morphologically the stems each stem was either:

1) fully recognised and classified to a morphological class (e.g. *f17*)
2) partially recognised and a set of morphological classes have been assigned
3) nothing was recognised

The explications above were regarding the third case. The first case is not interesting since the morphological class is already known and thus the unknown words are already classified. The second case however is not trivial. Observe that once we selected to concentrate on a specific stem, we know its gender and thus the gender of all word types it is supposed to cover. So, we do not have to care about the vector components that represent the other two genders.

There are three opportunities to deal with this:

a) just ignore the gender and proceed as in case 3) above

b) each time we have to deal with a predictor vector just clear its non-relevant gender components and renormalise to obtain conditional probability distribution given the gender

c) use other versions of the predictors vectors whose components contain conditional probabilities given the gender

We selected the last option c) and calculated additional conditional vectors for each gender. Obviously, this is much more correct especially what about the articles and the pronouns (because for the prepositions this is almost the same as if we applied b)).

When the whole pass through the text is performed the word types vector coordinates are normalised (divided by their sum) thus obtaining the word types vectors. Figure 34 shows the word type context vectors for the most frequent words for collection of 8.5 MB raw text files. Each word type is followed by the frequency of the useful contexts it has been found in. Then follows the distribution vector. The singular and the plural distribution are separated by tabulation. The three genders are output on different lines. These lines start with the first letter of the gender followed by the percents attributed to that gender.

```
Mann         599
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(46%)  25.00   0.20    2.00    15.00           0.03    1.70    1.90    0.01
     F(20%)  0.18    6.30    9.70    0.27            0.02    1.50    1.80    0.01
     N(34%)  11.00   0.12    1.20    19.00           0.02    1.40    1.50    0.01
Kind         351
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(7.1%) 6.50    0.00    0.36    0.08            0.05    0.02    0.03    0.03
     F(1.2%) 0.36    0.07    0.11    0.54            0.04    0.02    0.03    0.03
     N(92%)  47.00   0.00    0.27    45.00           0.03    0.02    0.02    0.02
Frau         334
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(12%)  3.70    0.98    0.00    1.10            3.30    0.90    0.19    1.90
     F(76%)  23.00   4.60    8.10    35.00           2.50    0.81    0.18    1.90
     N(12%)  3.60    0.60    0.00    3.20            2.10    0.75    0.15    1.60
Menschen             307
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(56%)  2.10    12.00   11.00   26.00           1.60    0.33    2.20    0.91
     F(20%)  5.20    1.30    2.40    6.40            1.20    0.30    2.00    0.91
     N(24%)  2.60    7.50    7.70    2.70            1.00    0.28    1.70    0.74
Mensch               287
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(36%)  34.00   0.00    0.00    0.00            0.00    2.00    0.00    0.00
     F(20%)  0.00    7.30    11.00   0.00            0.00    1.80    0.00    0.00
     N(44%)  16.00   0.00    0.00    26.00           0.00    1.60    0.00    0.00
Wort         278
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(13%)  13.00   0.00    0.21    0.00            0.04    0.01    0.00    0.02
     F(2.4%) 0.86    0.03    0.05    1.40            0.03    0.01    0.00    0.02
     N(84%)  42.00   0.00    0.21    42.00           0.02    0.01    0.00    0.02
Hand         274
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(16%)  1.50    1.20    3.00    1.40            5.40    0.18    0.18    3.10
     F(66%)  24.00   0.82    1.40    32.00           4.00    0.16    0.17    3.10
     N(18%)  4.80    0.71    1.70    4.70            3.50    0.15    0.14    2.50
Comanchen            262
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(29%)  13.00   0.00    0.62    4.50            4.10    3.40    1.30    2.30
     F(62%)  10.00   13.00   19.00   11.00           3.00    3.00    1.20    2.30
     N(9%)   0.00    0.00    0.62    0.00            2.70    2.80    1.00    1.80
Leben        260
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(14%)  5.90    0.27    4.10    2.90            0.12    0.16    0.09    0.10
     F(3.2%) 0.56    0.59    0.88    0.71            0.10    0.14    0.08    0.10
     N(83%)  41.00   0.17    2.50    39.00           0.08    0.13    0.07    0.08
Sache        258
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(10%)  2.10    0.58    0.00    0.00            4.50    0.53    0.00    2.60
     F(66%)  22.00   2.30    3.80    31.00           3.40    0.48    0.00    2.60
     N(24%)  9.40    0.35    0.00    8.40            2.90    0.44    0.00    2.10
Mädchen              252
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(17%)  8.80    0.00    6.90    0.24            0.57    0.11    0.10    0.31
     F(6.9%) 2.20    0.40    0.61    2.80            0.42    0.10    0.09    0.31
     N(76%)  35.00   0.00    5.00    35.00           0.37    0.09    0.08    0.25
Alte         246
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(25%)  17.00   0.00    0.00    0.00            2.00    4.60    0.00    1.10
     F(62%)  5.80    17.00   26.00   6.80            1.50    4.10    0.00    1.10
     N(13%)  3.90    0.00    0.00    3.40            1.30    3.80    0.00    0.88
Gesellschaft         230
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
     M(21%)  7.70    0.00    0.00    4.00            4.90    1.80    0.12    2.80
     F(70%)  20.00   6.50    9.90    26.00           3.60    1.60    0.11    2.80
     N(8.5%) 0.53    0.00    0.00    1.00            3.10    1.50    0.10    2.20
Tag          213
             NOM     GEN     DAT     AKK             NOM     GEN     DAT     AKK
```

M(56%)	11.00	0.00	0.00	38.00	0.21	1.10	5.50	0.11
F(19%)	0.80	4.10	6.60	1.10	0.15	0.99	5.10	0.11
N(25%)	8.60	0.00	0.00	11.00	0.13	0.92	4.30	0.09

Figure 34. Word type context vectors for the most frequent words from a raw text.

We can see that words like *der Mann* that are masculine have only 46% probability to be judged as such. This may be regarded as counter evidence that these context vectors can be useful. Though, we would like to stress that they are built exploiting context information only. In plus only a very limited amount of sources: just articles and combination of preposition followed by an article. These vectors will not be used as such but the conditional probability distributions will be used instead. This will be explained in more details below.

nom/sg	‘0’	4189						
	NOM	GEN	DAT	**AKK**	NOM	GEN	DAT	AKK
M(59.0076%)	**11.88**	0.72	2.85	**37.89**	0.24	1.38	3.91	0.13
F(20.5183%)	**0.87**	5.12	8.23	**1.12**	0.18	1.24	3.63	0.13
N(20.4741%)	**5.53**	0.44	1.77	**8.20**	0.15	1.15	3.12	0.11
gen/sg	‘[e]s(1)’		451					
	NOM	**GEN**	DAT	AKK	NOM	GEN	DAT	AKK
M(53.4709%)	0.66	**49.11**	0.71	1.30	0.87	0.15	0.19	0.48
F(9.30669%)	2.87	**0.57**	0.86	3.57	0.64	0.14	0.17	0.48
N(37.2224%)	2.81	**30.39**	0.41	2.39	0.56	0.13	0.15	0.39
dat/sg	‘[e]’	5264						
	NOM	GEN	**DAT**	**AKK**	NOM	GEN	DAT	AKK
M(52.6451%)	**10.50**	0.66	**4.18**	**30.88**	1.27	1.25	3.17	0.74
F(25.8516%)	**3.69**	4.66	**7.48**	**4.27**	0.94	1.12	2.94	0.74
N(21.5033%)	**5.61**	0.40	**2.67**	**7.84**	0.81	1.04	2.53	0.60
akk/sg	‘0’	4189						
	NOM	GEN	DAT	**AKK**	NOM	GEN	DAT	AKK
M(59.0076%)	**11.88**	0.72	2.85	**37.89**	0.24	1.38	3.91	0.13
F(20.5183%)	**0.87**	5.12	8.23	**1.12**	0.18	1.24	3.63	0.13
N(20.4741%)	**5.53**	0.44	1.77	**8.20**	0.15	1.15	3.12	0.11
nom/pl	‘”er’	1075						
	NOM	GEN	DAT	AKK	**NOM**	**GEN**	DAT	**AKK**
M(27.8381%)	5.13	0.43	9.34	3.54	**5.27**	**0.73**	0.28	**3.12**
F(46.6455%)	14.70	2.88	4.57	16.53	**3.93**	**0.66**	0.26	**3.12**
N(25.5163%)	5.91	0.27	6.21	6.40	**3.38**	**0.61**	0.22	**2.53**
gen/pl	‘”er’	1075						
	NOM	GEN	DAT	AKK	**NOM**	**GEN**	DAT	**AKK**
M(27.8381%)	5.13	0.43	9.34	3.54	**5.27**	**0.73**	0.28	**3.12**
F(46.6455%)	14.70	2.88	4.57	16.53	**3.93**	**0.66**	0.26	**3.12**
N(25.5163%)	5.91	0.27	6.21	6.40	**3.38**	**0.61**	0.22	**2.53**
dat/pl	‘”ern’	96						
	NOM	GEN	DAT	AKK	NOM	GEN	**DAT**	AKK
M(41.524%)	5.94	1.01	4.13	20.21	1.84	0.40	**6.97**	1.02
F(22.5247%)	4.53	1.49	2.37	4.86	1.36	0.36	**6.53**	1.02
N(35.9513%)	11.48	0.61	3.67	12.37	1.19	0.34	**5.47**	0.81
akk/pl	‘”er’	1075						
	NOM	GEN	DAT	AKK	**NOM**	**GEN**	DAT	**AKK**
M(27.8381%)	5.13	0.43	9.34	3.54	**5.27**	**0.73**	0.28	**3.12**
F(46.6455%)	14.70	2.88	4.57	16.53	**3.93**	**0.66**	0.26	**3.12**
N(25.5163%)	5.91	0.27	6.21	6.40	**3.38**	**0.61**	0.22	**2.53**

Figure 35. Context vectors for the class *m1*.

After the word types have their vectors built, we are ready to find the most appropriate class for each vector. During the training phase each inflexion class had its vector set: one vector for each possible ending. These vectors are easily obtained by the internal class distribution. The number of the vectors may vary for the different inflexion classes. Each vector is also assigned weight proportional to the conditional probability of occurrence of that ending given the inflexion class.

To get idea how the class vectors look like we present the context vector for *m1* on Figure 35. We estimated the *class* vectors as the average of the word type context vectors for the word forms that have first the specified class and then an acceptable ending given the class. The class *m1* has 5 different vectors since some of the forms are repeated more than once. The vectors for nom/pl, gen/pl and akk/pl are the same since the class endings for these forms are the same. For the same reasons the vectors for nom/sg and akk/sg are the same. An interesting case is the vector for dat/sg. Since it contains an optional *e* as ending it includes the zero ending word types as well. Thus, all the word types that contributed to build the vector for nom/sg and akk/sg are included for dat/sg as well. The reason to calculate the class context vectors looking at the ending only is that this is the information we have given a word form. We cannot choose between the different possibilities since we have no other information than the context and the ending. When we consider a particular stem and morphological class hypothesis we can conclude what the ending must be.

How the inflexion class probability given a specific stem is calculated? For each inflexion class among the feasible ones for the stem we calculate the possibly weighted sum of the dot products between the class vector and the corresponding unknown word type vector. Each cosine could be weighted accordingly (see above). This sum is then multiplied by the probability of that inflexion class given the stem. The obtained value is the inflexion class score given the stem. After the scores for all feasible inflexion classes are obtained they are normalised to get a kind of probability distribution. This distribution is output as the final product of the classification. The purpose of the cosine here is to compare two different distributions. We consider some other similarity measures for distributions comparison like KL divergence.

9 Future work

9.1 Short term

9.1.1 NEGRA nouns Stem Lexicon development

A crucial resource for the System is the Stem Lexicon, which is used for the basic model parameter estimation. This lexicon is currently automatically induced from the Morphy Lexicon. The Stem Lexicon created in that manner is not checked in its entirety and is error-prone. Prof. Walter von Hahn from the Hamburg University has created a special annotation tool to assign morphological classes to the nouns in the NEGRA corpus manually. This is done already for the nouns starting with the letter "*A*" and the morphological classes have been checked against the Stem Lexicon automatically induced from the Morphy Lexicon classes. There are only a very limited number (3 words) of words for which the two annotations disagree mainly for words that can have more than one class. On the other hand the Morphy Lexicon covers only part of the nouns found the NEGRA corpus (mainly due to compounds and foreign words), which harms the contextual parameters estimation. Thus, a manually annotated Stem Lexicon for the nouns in the NEGRA corpus is expected to help a lot in the model parameters estimation. The annotation of the nouns in the NEGRA corpus is important because we want to test and evaluate the model against the NEGRA corpus.

9.1.2 Model evaluation

The method will be evaluated against the NEGRA corpus. The corpus words will be separated in 10 groups of almost equal size on a random principle. Then a leave-one-out strategy will be used 10 times. The words from each of the groups will be removed from both the Stem Lexicon and Word Lexicon and held out. Then the algorithm will be started and its output on each step will be checked against the held out words and their manual annotation. The system will be evaluated in terms of precision, recall and coverage. This evaluation will reveal the steps where the System is most error-prone and allow us to concentrate especially there.

Several further refinement steps are under consideration currently but we are willing to apply them only after the complete System evaluation. We will then test whether and if yes, to which extent the refinement really improves System's performance for the particular step.

9.1.3 Model tuning

The model tuning will be applied only after the complete System evaluation against the NEGRA corpus using the manually annotated Stem Lexicon containing all the corpus nouns. Anyway, we have some ideas about how to tune or improve some specific details, which we believe could help to leverage the System performance. We present here a list of the most important ones. This list is not extensive and additional ideas may be added after the System evaluation is complete.

1. Better heuristic for the end of sentence as described in (Manning & Schuetze, 1999; Mikheev, 1999, 2000).

2. Comparison whether the statistics must be estimated on the rare words in the corpus or on all words.

3. Local context predictor's information that is used in the vector construction could be used earlier in the stem coverage.

4. The present model leaks any semantic information. Local semantic vectors could be created and used to test whether two words can share a common stem. A useful model using Latent Semantic Indexing is under consideration. (Schone & Jurafsky, 2000)

5. Collocations identification for better heuristic about "what is a noun?" If a capitalised word is met in the middle of a sentence but as a part of collocation than we cannot be sure why it is capitalised: because is a noun or because is a part of collocation (e.g. *neue* in *Neue Maxhütte*). We would prefer to treat it the same way as the words at the beginning of a sentence.

6. Account for the cases when a word type can have more than one stem.

7. Mikheev-style ending guessing rules predicting a set of 2, 3, 4 and 5 morphological classes.

8. More sophisticated treatment of the compounds: using the internal highly predictive endings as it has been done by (Adda-Decker and Adda, 2000).

9. Alternative approach to compounds. Using the longest matching sub-string approach, proposed by (Neumann and Mazzini, 1999; Neumann et al., 1997).

10. Use of sum of the morphological class probabilities instead of selecting the best single class when trying to impose maximum stem coverage principle.

11. An alternative heuristic for noun discovery: If a word is met capitalised in the middle of a sentence it is considered as a potential noun and is rejected otherwise.

12. Account for the German orthographic reform. This seems quite easy: we have just to convert all umlauts to their new two-letter equivalents as well as substitute *ß* with *ss*. We have to be careful about the "three consonants rules" as well.

13. More sophisticated classification that will be able to automatically induce the appropriate morphological classes in case a stem could have more than one morphological class *from the same gender*.

14. Design of special classes for the words that are used in either singular or plural form but not both.

15. Both deterministic and pr obabilistic context modules have to be linked as part of the System.

9.2 Long term

9.2.1 Application to other open-class POS

A similar approach could be applied to other important open-class POS such as: adjectives, verbs and adverbs. Obviously, this will not be straightforward but most of the steps (except perhaps the identification step) could be applied almost without any changes. Of course, special morphological classes for each distinct POS have to be defined as well as a stem lexicon in order to be able to estimate the model parameters (especially for the ending-guessing rules, as well as the different maximum likelihood estimates). The hardest thing there will be the automatic discovery of the specific POS instances since they will be non-capitalised and thus the heuristic used here will be unusable. A very promising approach could be to try to guess the POS of an unknown word using (Brill, 1995) or (Mikheev, 1997) style morphological and ending-guessing rules to find the POS of an unknown word. In fact we prefer to use the Mikheev's approach since it uses only a lexicon while Brill's approach relies on a tagged corpus, which is much harder to find.

9.2.2 Application to Bulgarian and Russian

The approach used here is not limited to German and could be applied to any inflectional language. In fact the more inflectional the language the better results are expected. That is why Bulgarian and Russian are good candidates. The very first thing to try in this direction is the application for Bulgarian nouns since the set of the 72 morphological classes as well as a lexicon are defined and available already. In fact the main and the hardest thing for Bulgarian will be the automatic unknown nouns identification. It was much easier in German where the nouns are capitalised. The usage of Mikheev-style ending guessing rules could be particularly useful.

10 Acknowledgements

I am very grateful to my advisor Prof. Galia Angelova from the Linguistic Modelling Department, Central Laboratory for Parallel Processing, Bulgarian Academy of Sciences (BIS-21 Center of Excellence) for the valuable discussions, suggestions and the strong support during all the work.

I am very grateful to Prof. Walther von Hahn from the Natural Language Systems group at the Computer Science department of Hamburg University for his support and valuable suggestions as well as for the especially developed resources. I would like to thank also Ingo Schröder from the same group for the valuable discussions and suggestions. The visit to Hamburg University gave me the opportunity to implement, evaluate and discuss my ideas on the system development in a native German speaking environment.

And lastly, but not in the last place, I would like to thank my reviewer Prof. Plamen Mateev from the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, for the valuable discussions and suggestions, especially in the statistical aspects of the present work.

11 References

Adda-Decker M., Adda G. (2000) *Morphological decomposition for ASR in German.* Phonus 5, Institute of Phonetics, University of the Saarland, pp.129-143. (http://www.coli.uni-sb.de/phonetik/phonus/phonus5/Adda.pdf)

Adda G., Adda-Decker M., Gauvain J., Lamel L. (1997). *Text normalization and speech recognition in French*. Proc. 5 th Conf. on Speech Comm. and Techn. (Eurospeech'97), Rhodes. (http://citeseer.nj.nec.com/adda97text.html)

Adda-Decker M., Adda G., Lamel L. & Gauvain J. (1996). *Developments in large vocabulary, continuous speech recognition of German.* Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'96), Atlanta.

Angelova G., Bontcheva K. (1996a) *DB-MAT: A NL-Based Interface to Domain Knowledge.* In Proceedings of the Conference "Artificial Intelligence - Methodology, Systems, Applications" (AIMSA-96), September 1996, Sozopol, Bulgaria. (http://lml.bas.bg/projects/dbr-mat/papers/iccs97/iccs97.html)

Angelova G., Bontcheva K. (1996b) *DB-MAT: Knowledge Acquisition, Processing and NL Generation using Conceptual Graphs.* In Proceedings of the 4th International Conference on Conceptual Structures (ICCS-96), August 1996, Sydney, Australia, LNAI, Springer-Verlag. (http://lml.bas.bg/projects/dbr-mat/papers/ranlp97/ranlp97.html)

Antworth E. (1990) *PC-KIMMO: a two-level processor for morphological analysis.* Occasional Publications in Academic Computing No. 16. Dallas, TX: Summer Institute of Linguistics. ISBN 0-88312-639-7, p. 273, paperbound.

Armstrong S., Russell G., Petitpierre D., Robert G. (1995) *An open architecture for multilingual text processing.* In: Proceedings of the ACL SIGDAT Workshop. From Texts to Tags: Issues in Multilingual Language Analysis, Dublin.

Bergenholtz H., Schaeder B. (1977). *Die Wortarten des Deutschen. Versuch einer syntaktisch orientierten Klassifikation.* Stuttgart: Klett, 243 p.

Brill E. (1999). *Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging;* In Natural Language Processing Using Very Large Corpora, 1999. (http://research.microsoft.com/~brill/Pubs/unsuprules.ps)

Brill E. (1995) *Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging.* In Computational Linguistics, 21(4):543-565. (http://research.microsoft.com/~brill/Pubs/recadvtagger.ps)

Cucerzan S., Yarowsky D. (2000) *Language independent minimally supervised induction of lexical probabilities.* Proceedings of ACL-2000, Hong Kong, pages 270-277, 2000. (http://www.cs.jhu.edu/~yarowsky/pdfpubs/acl2000_cy.ps)

Cutting, Doug, Kupiec J., Pedersen J., Sibun P. (1992) *A practical part-of-speech tagger.* Proceedings of the Third Conference on Applied Natural Language Processing (ANLP-92), pp. 133-140, 1992. (http://citeseer.nj.nec.com/cutting92practical.html)

Daciuk J. (1997) *Treatment of Unknown Words.* (http://citeseer.nj.nec.com/354810.html)

Daciuk, J., Watson R, and Watson B. (1998) *Incremental construction of acyclic finite-state automata and transducers.* In Finite State Methods in Natural Language Processing, Bilkent University, Ankara, Turkey. (http://citeseer.nj.nec.com/337966.html)

DeJean H. (1998) *Morphemes as necessary concepts for structures: Discovery from untagged corpora.* University of Caen-Basse Normandie. (http://www.info.unicaen.fr/~DeJean/travail/articles/pg11.htm)

Deshler D., Ellis E., Lenz B. (1996) *Teaching Adolescents with Learning Disabilities: Strategies and Methods.* Love Publishing Company, 1996.

Dietmar E. and Walter H. (1987) *Bulgarisch-Deutsch Wörterbuch.* VEB Verlag Enzyklopädie Leipzig, 1987

Drosdowski G. (1984). *Duden. Grammatik der deutschen Gegenwartssprache.* Dudenverlag, Mannheim.

Finkler W., Lutzky O. (1996) *MORPHIX.* In Hausser, R. (Ed.): Linguistische Verifikation. Dokumentation zur ersten Morpholympics 1994. Tübingen: Niemeyer, pp. 67-88, 1996.

Finkler W., Neumann G. (1988) *MORPHIX. A Fast Realization of a Classification-Based Approach to Morphology.* In: Trost, H. (ed.): 4. Osterreichische Artificial-Intelligence-Tagung. Wiener Workshop - Wissensbasierte Sprachverarbeitung. Proceedings. Berlin etc. pp. 11-19, Springer, 1988. (http://www.dfki.de/~neumann/publications/new-ps/morphix88.ps.gz)

Finkler W., Neumann G. (1986) *MORPHIX - Ein hochportabler Lemmatisierungsmodul fur das Deutsche.* FB Informatik, KI-Labor, Memo Nr. 8, Juli 1986.

Gaussier E. (1999) *Unsuppervised learning of derivational morphology from inflectional lexicons.* ACL'99 Workshop Proceedings: Unsupervised Learning in Natural Language Processing., University of Maryland, 1999. (http://www.xrce.xerox.com/publis/mltt/gaussier-egulnlp-99.ps)

Goldsmith J. (2000) *Unsupervised Learning of the Morphology of a Natural Language.* Version of April 25, 2000. To appear in Computational Linguistics (2001). (http://humanities.uchicago.edu/faculty/goldsmith)

Goldsmith J., Reutter T. (1998) *Automatic collection and analysis of German compounds.* In The Computational Treatment of Nominals: Proceedings of the Workshop COLING-ACL '98. Montreal. Edited by Frederica Busa, Inderjeet Mani and Patrick Saint-Dizier. pp. 61-69. 1998.

Haapalainen M., Majorin A. (1994) *GERTWOL: Ein System zur automatischen Wortformerkennung Deutscher Wörter.* Lingsoft, Inc., September 1994. (http://www.ifi.unizh.ch/CL/gschneid/LexMorphVorl/Lexikon04.Gertwol.html)

Hafer M, Weiss S. (1974) *Word segmentation by letter successor varieties.* Information Storage and Retrieval, 10.

Harman, D. (1991) *How effective is suffixing?* In Journal of The American Society of Information Science. Vol. 42, No 1. 1991.

Hietsch, O. (1984). *Productive second elements in nominal compounds: The matching of English and German.* Linguistica 24, pp. 391-414.

Hoch R. (1994) *Using IR Techniques for Text Classification in Document Analysis.* In: Proceedings of 17th International Conference on Research and Development in Information Retrieval (SIGIR'94), Dublin City, Ireland, 1994. (http://citeseer.nj.nec.com/hoch94using.html)

Hull, D. (1996) *Stemming Algorithms: A Case study for detailed evaluation.* In Journal of The American Society of Information Science. Vol. 47, No 1. 1996.

Jacquemin, C. (1997) *Guessing morphology from terms and corpora.* In Actes, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97), pp. 156–167, Philadelphia, PA.

Karp D., Schabes Y., Zaidel M., Egedi D. (1992) *A freely available wide coverage mophological analyzer for English.* In: Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France, 1992. (http://citeseer.nj.nec.com/daniel92freely.html)

Karttunen L. (1983) *KIMMO: a general morphological processor.* Texas Linguistic Forum 22:163-186.

Kazakov D. (1997) *Unsupervised Learning of Na?ve Morphology with Genetic Algorithms.* In W. Daelemans, A. van den Bosch, and A. Weijtera, eds., Workshop Notes of the ECML/Mlnet Workshop on Empirical Learning of Natural Language Processing Tasks, April 26, 1997, Prague.

Koskenniemi K. (1993) *Glossing text with the PC-KIMMO morphological parser.* Computers and the Humanities 26:475-484.

Koskenniemi K. (1984) *A general computational model for word-form recognition and production.* In COLING 1984 pp. 178 – 181, Stanford University, California, 1984.

Koskenniemi, K. (1983a) *Two-level morphology: a general computational model for word-form recognition and production*. Publication No. 11. University of Helsinki: Department of General Linguistics.

Koskenniemi K. (1983b) *Two-level model for morphological analysis.* In IJCAI 1983 pp. 683-685, Karlsruhe, 1983.

Kraaij W. (1996) *Viewing stemming as recall enhancement.* In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York.1996. (http://citeseer.nj.nec.com/kraaij96viewing.html)

Krovetz R. (1993) *Viewing Morphology as an Inference Process.* Proceedings of the Sixteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval 1993: pp. 191-202. (http://citeseer.nj.nec.com/krovetz93viewing.html)

Kupiec J. (1992) *Robust part-of-speech tagging using a hidden Markov model.* Computer Speech and Language, 6(3), pp.225-242, 1992.

Lamel L., Adda-Decker M. & Gauvain J. (1995). *Issues in large vocabulary, multilingual speech recognition.* Proc. 4th Conf on Speech Comm. and Techn. (Eurospeech'95), Madrid. (http://citeseer.nj.nec.com/173875.html)

Lezius W. (2000) *Morphy - German Morphology, Part-of-Speech Tagging and Applications*. In Ulrich Heid; Stefan Evert; Egbert Lehmann and Christian Rohrer, editors, Proceedings of the 9th EURALEX International Congress pp. 619-623 Stuttgart, Germany. (http://www-psycho.uni-paderborn.de/lezius/paper/euralex2000.pdf)

Lezius W., Rapp R., Wettler M. (1998) *A Freely Available Morphological Analyzer, Disambiguator, and Context Sensitive Lemmatizer for German*. In Proceedings of the COLING-ACL 1998 pp. 743-747. (http://www-psycho.uni-paderborn.de/lezius/paper/coling.pdf)

Lezius W., Rapp R., Wettler M. (1996a) *A Morphology-System and Part-of-Speech Tagger for German.* In: D. Gibbon, editor, Natural Language Processing and Speech Technology. Results of the 3rd KONVENS Conference. pp. 369-378 Mouton de Gruyter. (http://www-psycho.uni-paderborn.de/lezius/paper/konvens.pdf)

Lezius W. (1996b). *Morphologiesystem MORPHY.* In: R. Hausser, ed., Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994, pp. 25-35. Niemeyer, Tübingen. (http://www-psycho.uni-paderborn.de/lezius/paper/molympic.pdf)

Lorenz O. (1996). *Automatische Wortformenerkennung für das Deutsche im Rahmen von Malaga*. Magisterarbeit. Friedrich-Alexander-Universität Erlangen-Nürnberg, Abteilung für Computerlinguistik. (http://www.linguistik.uni-erlangen.de/tree/PS/dmm.ps)

Lovins J. (1968) *Development of a stemming algorithm*. Mech. Trans. And Comp. Ling. 11. 1968.

Manning C., Shuetze H. (1999) *Foundations of Statistical Language Processing.* MIT Press 1999 ISBN 0262133601. (http://nlp.stanford.edu/fsnlp/)

Matsuoka, T., Ohtsuki, K., Mori, T., Furui, S. & Shirai, K. (1996). *Large vocabulary continuous speech recognition using a Japanese business newspaper (Nikkei).* Proc. DARPA Speech Recognition Workshop, Harriman, pp. 137-142.

Mikheev A. (2000). *Tagging Sentence Boundaries.* In NACL'2000 (Seattle) ACL April 2000. pp. 264-271. (http://www.ltg.ed.ac.uk/~mikheev/papers_my/nacl_00.ps)

Mikheev, A (1999). *Periods, Capitalized Words, etc.* Computational Linguistics, 1999. pp. 25. (http://www.ltg.ed.ac.uk/~mikheev/papers_my/cl-prop.ps)

Mikheev A. (1997). *Automatic Rule Induction for Unknown Word Guessing.* In Computational Linguistics vol 23(3), ACL 1997. pp. 405-423. (http://www.ltg.ed.ac.uk/~mikheev/papers_my/cl-unknown.ps)

Mikheev A. (1996a). *Learning Part-of-Speech Guessing Rules from Lexicon: Extension to Non-Concatenative Operations.* In Proceedings of the 16th International Conference on Computational Linguistics (COLING'96) University of Copenhagen, Copenhagen, Denmark. August 1996. pp. 237-234. (http://www.ltg.ed.ac.uk/~mikheev/papers_my/col-96.ps)

Mikheev A. (1996b). *Unsupervised Learning of Part-of-Speech Guessing Rules.* In Journal for Natural Language Engineering. vol 2(2). Cambridge University Press. 1996. (http://www.ltg.ed.ac.uk/~mikheev/papers_my/jnlp-unknown.ps)

Mikheev A. (1996c). *Unsupervised Learning of Word-Category Guessing Rules.* Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, University of California, Santa Cruz, pp. 62-70, 1996. (http://www.ltg.ed.ac.uk/~mikheev/papers_my/acl-96.ps)

Neumann G., Mazzini G. (1999) *Domain-adaptive Information Extraction.* DFKI, Technical Report, 1999. (http://www.dfki.de/~neumann/smes/smes.ps.gz)

Neumann G., Backofen R., Baur J., Becker M., Braun C. (1997) *An Information Extraction Core System for Real World German Text Processing.* In Proceedings of 5th ANLP, Washington, March, 1997. (http://www.dfki.de/cl/papers/cl-abstracts.html#smes-anlp97.abstract)

Petitpierre D., Russell G. (1995) *MMORPH - the Multext morphology program.* Technical report, ISSCO, 54 route des Acacias, CH-1227 Carouge, Switzerland, October 1995.

Popovic M., Willett P. (1992) *The Effectiveness of Stemming for Natural Language access to Slovene Textual Data.* In Journal of The American Society of Information Science. Vol. 43, No 5. 1992.

Porter M. (1980) *An algorithm for suffix stripping.* Program 14, 3. 1980. (http://telemat.det.unifi.it/book/2001/wchange/download/stem_porter.html)

Rapp R. (1996) *Die Berechnung von Assoziationen. Ein korpuslinguistischer Ansatz.* Olms, Hildenheim, 1996. (http://www.fask.uni-mainz.de/user/rapp/papers/disshtml/main/main.html)

Rostek L., Alexa M. (1998). *Marking up in TATOE and exporting to SGML: Rule development for identifying NITF categories.* In Computers and the Humanities, Vol. 31/4, 1998. (http://www.cs.queensu.ca/achallc97/papers/p029.html)

Schmid H. (1995). *Improvements in part-of-speech tagging with an application to German*. In: Feldweg and Hinrichs, eds., Lexikon und Text, pp. 47-50. Niemeyer, Tübingen. (http://citeseer.nj.nec.com/schmid95improvement.html)

Schone P., Jurafsky D. (2000) *Knowledge-Free Induction of Morphology Using Latent Semantic Analysis.* In Proceedings of CoNLL-2000 and LLL-2000, pp. 67-72, Lisbon, Portugal, 2000. (http://lcg-www.uia.ac.be/conll2000/abstracts/06772sch.html)

Sproat R. (1991) *Review of "PC-KIMMO: a two-level processor for morphological analysis"* by Evan L. Antworth. Computational Linguistics 17.2:229-231.

Thede S., Harper M. (1997) *Analysis of Unknown Lexical Items using Morphological and Syntactic Information with the TIMIT Corpus.* Proceedings of the Fifth Workshop on Very Large Corpora, August 1997. (http://citeseer.nj.nec.com/thede97analysis.html)

Thede S. (1997) *Tagging Unknown Words using Statistical Methods.* (http://citeseer.nj.nec.com/14497.html)

Trost, H. (1991) *X2MORF: A Morphological Component Based on Augmented Two-Level Morphology*. InProceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91). Sydney, Australia.

Trost H., Dorffner G. (1985) *A system for morphological analysis and synthesis of German texts.* In D.Hainline, editor, Foreign Language CAI. Croom Helm, London, 1985.

Ulmann, M. (1995) *Decomposing German Compound Nouns*. Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria, 265-270.

v. Hahn W., Angelova G. (1996) *Combining Terminology, Lexical Semantics and Knowledge Representation in Machine Aided Translation.* In: TKE'96: Terminology and Knowledge Engineering. Proceedings of the Conference "Terminology and Knowledge Engineering", August 1996, Vienna, Austria. pp. 304 – 314. (http://nats-www.informatik.uni-hamburg.de/~dbrmat/abstracts/tke96.html)

v. Hahn W., Angelova G. (1994) *Providing Factual Information in MAT.* In: Proceedings of the Conference "MT - 10 Years on", Cranfield, UK, November 1994, pp. 11/1 - 11/16. (http://nats-www.informatik.uni-hamburg.de/~dbrmat/abstracts/MAT94.html)

Van den Bosch, A. and W. Daelemans. (1999) *Memory-based morphological analysis.* Proc. of the 37th Annual Meeting of the ACL, University of Maryland, pp. 285-292. (http://citeseer.nj.nec.com/221820.html)

Viegas E., Onyshkevych B., Raskin V. and Nirenburg S. (1996) *From Submit to Submitted via Submission: On Lexical Rules in Large-Scale Lexicon Acquisition*. In Proceedings ACL96, pp. 32-39. ACL, 1996.

Weischedel R., Meeter M., Schwartz R., Ramshaw L. and Palmucci J. (1993) *Coping with ambiguity and unknown words through probabilistic models.* Computational Linguistics, 19:359-382, 1993.

Xu J., Croft B. (1998) *Corpus Based Stemming Using Coocurrence of Word Variants.* In ACM Transactions on Information Systems, Vol. 16, No 1. 1998. (http://citeseer.nj.nec.com/32742.html)

Yarowsky D. Wicentowski R. (2000) *Minimally supervised morphological analysis by multimodal alignment.* Proceedings of ACL-2000, Hong Kong, pp. 207-216, 2000. (http://www.cs.jhu.edu/~yarowsky/pdfpubs/acl2000_yar.ps)

Young, S., Adda-Decker, M., Aubert, X., Dugast, C., Gauvain, J.-L., Kershaw, D., Lamel,, L., Leeuwen, D., Pye, D., Robinson, A., Steeneken, H. & Woodland, P. (1997). *Multilingual large vocabulary speech recognition: the European SQALE project.* Computer Speech and Language 11(1), pp. 73-89.

11.1 Useful Links

Morphologiesystem Morphy
http://www-psycho.uni-paderborn.de/lezius/

Tatoe — Corpus query tool that imports the Morphy output
http://www.darmstadt.gmd.de/~rostek/tatoe.htm

PC-KIMMO: A Two -level Processor for Morphological Analysis
http://www.sil.org/pckimmo/about_pc-kimmo.html

GERTWOL
http://www.lingsoft.fi/doc/gertwol/intro/overview.html

Cogilex QuickTag and QuickParse
http://www.cogilex.com/products.htm

Malaga: a System for Automatic Language Analysis
http://www.linguistik.uni-erlangen.de/~bjoern/Malaga.en.html

Deutsche Malaga -Morphologie
http://www.linguistik.uni-erlangen.de/~orlorenz/DMM/DMM.en.html

Morphix
http://www.dfki.de/~neumann/morphix/morphix.html

Finite state utilities by Jan Daciuk
http://www.pg.gda.pl/~jandac/fsa.html

Canoo.com — Morphological resources on the Web. Useful morphological browser available.
http://www.canoo.com/online/index.html

NEGRA corpus
http://www.coli.uni-sb.de/sfb378/negra-corpus/negra-corpus.html

Die Wortformen der geschlossenen Wortarten im Stuttgart-Tübingen Tagset (STTS)
http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/Wortlisten/WortFormen.html

Expanded Stuttgart-Tübingen Tagset (STTS)
http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html

DB-MAT project
http://nats-www.informatik.uni-hamburg.de/~dbrmat/db-mat.html

DBR-MAT project
http://lml.bas.bg/projects/dbr-mat/

Natural Language Software Registry
http://registry.dfki.de

European Corpus Initiative
http://www.coli.uni-sb.de/sfb378/negra-corpus/cd-info-e.html

Linguistic Data Consortium
http://www.ldc.upenn.edu/

11.2 Tables Index

11.3 Figures Index