Automatic Recognition and Morphological Classification of Unknown German Nouns

Preslav Nakov[1], Galia Angelova[2], Walther von Hahn[3]

The work presented here was performed 2001 as a scientific project of the BIS-21 "Center of Excellence" project, ICA1-2000-70016 and was supported by the cooperation between Hamburg University Sofia University "St. Kl. Ohridski"

Abstract

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

Zusammenfassung

Der Bericht beschreibt ein System zur Erkennung und morphologischen Klassifizierug von einem deutschen Lexikon unbekannten Wörtern. Das System **MorphoClass** liest deutschen Rohtext und gibt eine Liste der im Lexikon nicht verzeichneten Wörter mit Hypothesenmengen über Ihre Wortartzugehörigkeit, ihr grammatisches Geschlecht, ihren Stamm und ihre morphologische Klasse aus. **MorphoClass** nutzt dabei globale Regeln (Endungsschätzung, maximum likelihood estimations, Worthäufigkeitsinformation), lokale Regeln (Kotext) und linguistisches Wissen (Regeln über Zusammensetzung, Flexion, Ableitung)

MorphoClass ist ein Werkzeug zum Aufbau eines Lexikons. Zur Zeit werden nur nomen verarbeitet, aber es werden auch Vorschläge zur Erweiterung des Ansatzes und zur Verarbeitung anderer flektierender Sprachen diskutiert.

---

[1] Linguistic Modelling Department, Central Laboratory for Parallel Processing ,Bulgarian Academy of Sciences (CLPP-BAS)
[2] Linguistic Modelling Department, Central Laboratory for Parallel Processing ,Bulgarian Academy of Sciences (CLPP-BAS)
[3] University of Hamburg, Germany, Computer Science Department

CONTENTS

Index of Tables

TERMINOLOGY

PoS	Part-of-Speech
Word type	in Information Retrieval, group of tokens with exactly the same graphemic form.
Ending	The string following the stem for inflexion. Endings are represented in the table of the 39 German morphological classes (see page 13). In these tables, however, the notion is extended to account for the umlaut- and ß-alternations.
End String	a sequence of characters at the end of a word. This string may be an ending but could be any set which is statistcly significant for the assignment of PoS.
Base form	the singular nominative form of the German nouns.
Stem	in this work, the string shared by all inflected forms of a noun. The changes caused by umlauts and words ending by “ß” are not considered to change the stem although they actually do.

1 Introduction

1.1 The problem

The recognition and effective processing of unknown words in a given text is a primary problem for any Natural Language Processing (NLP) system. No matter how big the applied lexicon is, there will always remain unknown words. Moreover, natural language is dynamic and it is impossible to compile huge dictionaries, which will contain all the words that could appear in real-life texts: New words are constantly added to a language, other words get less frequent ore are dropped out, while some of the existing ones lose, change or obtain new PoS features, gender, meaning etc. Even if one would manage to build a complete dictionary at a time, it will be outdated in only few days since new words will inevitably appear.

In most natural languages, the two major and typical classes of new words are the *proper nouns* and the *foreign words*. They cause significant problems in NLP applications because they are uncontrollable and theoretically unlimited. It is rather unlikely to imagine that all the proper nouns (e.g. company names) can be listed in exhaustive dictionaries; it is impossible to know all the names of places, persons or companies all over the world. Similarly, nobody can predict all the foreign words that could enter the language.

Another important source of new words are morphological processes, which directly influence the lexical stock of a language. There are three major linguistic phenomena in this field: *inflexion*, *derivation,* and *compound*.

The **inflexion** is very unlikely to produce an unknown wordform 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 and predictable. The inflection rules may differ according to gender, final position, and the possibility of umlaut for the main vovel. But inflexion is quite stable and languages tends to have/develop a limited set of morphological classes that cover all wordforms; only very few exceptions are found. The inflexion of new words that enter a language most often follow these established inflexional patterns.

The **derivation** process, compared to inflexion, is theoretically more powerful with respect to production of unknown words. Unlike inflexion, derivation produces words that – in most cases - have different *part of speech* (PoS) features. A word obtained through derivation is a new word and not just another form of the stem. The words obtained through derivation would be listed in a paper dictionary as separate entries while the inflected word forms are not presented there, each. But linguistics found out that the derivation patterns (suffixes, prefixes and infixes) are also well-established and in this way 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 amounts of word forms. The power of these processes differs from language to language. Slavonic and Roman languages, e.g., are highly inflexional while English is poor in inflexion, and poor in derivation, too, if compared to e.g. Bulgarian. However, even in English a considerable amount of words are produced by inflexion, while this is not the case with compounding.

Compounding is the process of direct concatenation of two or more words to form a new word with, possibly, a non-compositional meaning. Only few European languages use compounds to a larger extent, but especially for German it is an important factor. The compounding process is very powerful in German since it is fully productive. A German word can be part of a nearly unlimited amount of different compounds with other words. The process is not only theoretically very important, but also in practice: a large part of the unknown words in German stem from compound production.

There are other important sources of **pseudo-unknown** words in real texts: (1) misspelled words. (2) especially for German there is another recent source of new word forms: the orthographic reform, which is yet not completely accepted. Since some part of the population keeps the old orthography (even one of the biggest daily newspapers) and others use the new rules, this results in a variety of new word forms.

1.2 The system: MorphoClass

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

MorphoClass 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 word forms of the base while the morphological class describes both the word gender and the inflexion pattern for changes by case and number. The stem and the morphological class together determine all wordforms that could be obtained through inflexion in an unambiguous way.

MorphoClass solves the task in a sequence of subtasks including:

- unknown word identification,
- noun identification,
- recognition and grouping of inflected forms of the same word (i.e.sharing the same stem),
- compound splitting,
- morphological stem analysis,
- stem hypotheses for each group of inflected forms, and, finally,
- production of a ranked list of hypotheses about a possible morphological class for each group of words.

This is a complex multi-stage process, which exploits:

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

The current version of **MorphoClass** does not yet contain modules for processing local context.

1.3 Areas of application

What is **MorphoClass** and what is not? It 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 by new words. In the following rather short overview we set off **MorphoClass′ function against other similar linguistic tools.**

1.3.1 PoS guesser

MorphoClass is not a PoS guesser in a traditional meaning. The purpose of a PoS guesser is to make hypotheses about possible PoSs for an unknown word looking at its graphemic form in the particular local context and possibly in a lexicon. **MorphoClass** is not restricted to this local word context ; it collects and considers all the word occurrences throughout a complete input text. Moreover, we are not interested in the exact PoS of a word but just in whether it is a noun. Once we know that it is a noun, in contrast to PoS guessers, **MorphoClass** continues work by trying to identify other inflectional forms of the same word and derives a hypothesis for its morphological class (this includes gender identification). In this way, **MorphoClass** could be seen as kind of a **morphological class guesser**, which for instance might work after a PoS-tagger completed its task and taged the unknown nouns.

1.3.2 Morphological analyser

MorphoClass is not a pure morphological analyser although it can be used as such, since in the end it outputs the morphological information available for the known words just like a morphological analyser. However, it works at a global level, which means it does not try to disambiguate between possible lexical forms of a specific word token. **MorphoClass** is not interested in a particular word token in a specific context but in the word type, which the token is instance of. Morphological analysers usually list all possible morphological information and sometimes try to disambiguate possible morphological formsof the instance. In the latter 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. Morphological analysers usually apply some local strategies to deal with unknown words but it is not a central task for them and they often use simple heuristics only. Thus, **MorphoClass** can be looked at as a guessing extension of a morphological analyser.

1.3.3 Stemmer

MorphoClass is not a stemmer in the classic sense, although it outputs the stem 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* and *generator* would be grouped together by a classic stemmer but not by **MorphoClass**.

1.3.4 Lemmatiser

MorphoClass 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 defined in the morphological classes. 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 **MorphoClass** although this is not the central task. 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 possibilities. But we never output the splitting of the first part, although we always obtain it internally.

2 Related work

The **MorphoClass** task is more or less related to several classical NLP tasks: the nearest one being the morphological analysis, while other tasks like stemming are much more dissimilar. Below we discuss briefly related work.

(Deshler, Ellis & Lenz, 1996) present 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 is the basis for 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) developed *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. A similar approach is presented in (Schmid, 1995). (Rapp, 1996) derived 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 checks 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). This approach is discussed in more detail in section 3 below. 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 approach 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).

3 Ending-Guessing Rules

While most of the rules generated through a 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 the construction of automatic ending-guessing rules. We implemented a Mikheev-like ending-guessing rules mechanism(Mikheev, 1997).

Mikheev originally proposed it for POS guessing; we applied the same approach for morphological class guessing. We selected a 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 words against the Expanded Stem Lexicon and from there against the Stem Lexicon. We keep only rules with confidence score of at least 0.90 and frequency of 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: German literature and Reuters news. The list of all present endings is given in Appendix 1.

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

- *unambiguous* (predicts a particular class without or with only few exceptions. The fewer the exceptions, the better is 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.), and
- *long* (the ending length is another important argument. The longer the ending, the less is the probability that it will appear by chance, and thus the better is its prediction.).

What we need is a score for the rules that takes into account at least these three criteria (and maybe more). Doubtlessly, the most important factor is the rule ambiguity. We prefer rules which are 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. A straight-forward way to do so is to investigate the stems from the Stem Lexicon: to 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. However, this is not a very good idea since the words, which the stems represent are not equally likely in a real text. It is much better to estimate the frequencies x and n in 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 in 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 addition, 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 as possible of real language. Usually, a large text collection is used mostly from newspapers since they are supposed to be very representative of contemporary language and to cover a large variety of different fields.

So, we saw that, although maximum likelihood estimation is a good predictor of the rule accuracy, it is less useful for 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* π, 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:

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

where p is a modified version of $\hat{p}$ that ensures neither p nor $(1-p)$ could be zero : $p=(x+0.5)/(n+1)$; $\sqrt{\frac{p(1-p)}{n}}$ is an estimation of the dispersion; and $t_{(1-\alpha)/2}^{(n-1)}$ is a coefficient of the *t*-distribution.

The *t*-distribution $t_{(1-\alpha)/2}^{d}$ has two parameters: the degree of freedom d and the confidence level.

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 parameters being equal. (Mikheev, 1997) proposes to use the logarithm of the ending length l in a score of the form:

$$score = p - \frac{t_{(1-\alpha)/2}^{(n-1)} \sqrt{\frac{p(1-p)}{n}}}{1+\log(l)}, \; p = (x+0.5)/(n+1)$$

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 but we use 0.90 in order to obtain rules of higher quality (although less in number).

The ending-guessing have been estimated twice: once directly from the Stem Lexicon and once from a raw text collection. Tables 1 and 2 show the top members of the ending-guessing rules of the ranked rules list. MorphoClass system currently uses 1789 ending-guessing rules obtained from lexicons as well as raw text estimation (see Appendix 1).

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

Table 1.**Top ending guessing rules (lexicon).**

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

Table 2. Top ending guessing rules (raw text)

4 Morphological classes

The morphological classification in **MorphoClass** follows the one developed under the DB-MAT and DBR-MAT projects, which was anelaboration of the classification presented in (Dietmar and Walter, 1987). DB-MAT was 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 project was an extension of DB-MAT with a new language: Romanian. In these two projects, German morphology was necessary for the generator presenting German subject information to the user (Angelova & Bontcheva, 1996a, 1996b), as well as for the acquisition of the German lexicon with especially developed tool for lexicon acquisition (v.Hahn, 1999, 2002). More information about DB(R)-MAT could be found at http://nats-www.informatik.uni-hamburg.de/~dbrmat/ and http://www.lml.bas.bg/projects/dbr-mat/.

MorphoClass works with 41 inflexional classes for German nouns which were practically reduced to 39 since distinguishing the stress alternations is not important for MorphoClass (see Table 3, the inflexional classes *m9a* and *f16a* are considered as equivalent to *m9* and *f16* correspondingly). For convinience, the inflexional classes are marked according to the gender: there are 14 classes for masculine nouns (*m1-m11*), 10 classes for feminine nouns (*f12-f19*) and 15 classes for neutrum nouns (*n20-n31*).

4.1 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 by “s / ß / sch / x / chs / z / tz” the vowel “e” is obligatory.
b) when “ß” comes after a short vowel in the basic form, it is written as “ss” in all forms of the paradigm (old orthography).
2 – concerns the suffix in “dat pl”, “masc/neut” and means: if the basic form ends by “n” there is no second “n” as “dat pl” suffix.

Note:
Rule 1a), as an example, is not obligatory. It is just a preference for modern German; the bracketed form represents an older historic layer. 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 for 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	Bekan-nte
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
~~*m9*~~	~~*0*~~	~~*s*~~	~~*0*~~	~~*0*~~	~~*en*~~	~~*en*~~	~~*en*~~	~~*en*~~	~~*Direktor*~~

~~**a**~~									
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. *DB-MAT* morphological classes, corresponding alternation rules and sample 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	Fir-men	Fir-men	Fir-men	Fir-men
Floß/**n20a**	Floß	Flosses	Floß Flosse	Floß	Flösse	Flösse	Flössen	Flösse

Table 4. Example: Inflexion of German nouns by application of the rules of the corresponding morphological class.

Note:
The classes *n24* and *m6* in Table 3 have absolutely identical endings, so they are equivalently probable for a guesser of the morphological class. In this way the ambiguity of the inflexional endings is the first reason for losing precision in **MorphoClass**.

5 Resources used

Figure 5 shows the two kinds of German corpus/lexicon resources we used in addition to the morphological classes of German nouns.

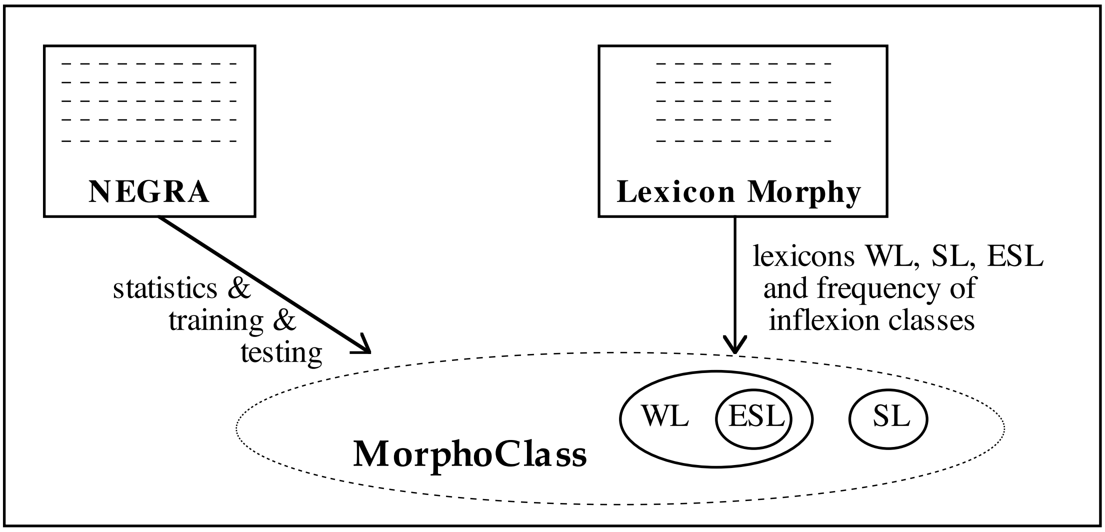

Table 5: Lexical resources

5.1 Morphologically annotated corpus NEGRA

This study used the NEGRA corpus which consists of approximately 176,000 tokens (10,000 sentences) of German newspaper text. The first 60,000 tokens are not only PoS annotated but also morphologically annotated using the expanded Stuttgart-Tubingen-Tagset. We used this corpus to derive some statistics during the system's training phase, as well as to test MorphoClass' performance. For instance, the unknown stems in Table 4 and Table 5 came from NEGRA corpus.

5.2 Lexicons

We used the free lexicon of the morphological system Morphy by Lezius to automatically generate **MorphoClass** lexicons: *Word Lexicon* (*WL*), *Stem Lexicon* (*SL*) and *Expanded Stem Lexicon* (*ESL*). The original Morphy fullform lexicon contained 50,597 stems (17380 nouns, 22184 adjectives, 1409 proper nouns etc.) and 324,000 different wordforms. 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, morphological class>

that could generate these forms.

In this way, the morphological classes shown in Table 3 have been induced automatically from the Morphy wordforms and their corresponding morphological tags in Morphy lexicon. For each morphological class successfully induced, a stem was determined and recorded in SL. Additionally, all wordforms of the stem were recorded in the Expanded Stem Lexicon. Some Morphy wordforms belong to defected paradigms (e.g. only the wordforms in singular exist), no stem could be "calculated" for them and they were not recorded in the ESL.

5.2.1 Word Lexicon

We use a *Word Lexicon* containing 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. The lexicon does not necessarily contain all the inflected forms of a word and 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. For instance, each word in the beginning of a sentence is always written capitalised *regardless* of its PoS. On the other hand not all capitalised words in a non-starting position in a sentence can be unconditionally considered as nouns. (In the 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.

• 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.

5.2.2 Stem Lexicon

The *Stem Lexicon* contains a list of the known stems together with their morphological class. The Stem Lexicon currently contains 13,072 stems with 13,147 different classes (*der/die/das Halfter* has three different morphological classes: *m4*, *f16* and *n23*, and 73 other stems having two different morphological classes). These 13072 stems were automatically produced from the 17380 stems of nouns in the Morphy lexicon. All "defected" stems from Morphy, which were not successfully treated by our algorithm for recognition of stem, were omited in **MorphoClass**.

5.2.3 Expanded Stem Lexicon

The *Expanded Stem Lexicon* is "the cover" of the Stem Lexicon. It contains a generated list of all full paradigms of the correctly recognised 13072 stems in SL. Usually, there are 8 wordforms per stem, one form per case/number combination, but sometimes the wordforms could be 9 or 10 since some of the rules have optional elements (especially in gen/sg) and produce dublets. The classes *m1a*, *m7*, *n20a*, *n26*, *n27*, *n31* have one optional element and thus produce 9 forms, and *m1*, *m2*, *m3*, *m3a*, *m9*, *n20*, *n21*, *n22*, *n25* produce 10 forms per stem.

What is really important is that the Expanded Stem Lexicon contains *all* the forms whose stems are known as nouns. The same applies to the Word Lexicon: it contains ESL and all other known words. We rely on these properties - while recognising unknown nouns and guessing their stems - to reject the known stems as candidates for the unknown words' stems. An *unknown* word cannot have a known stem since all the words that have this stem are supposed to be included in the Expanded Stem Lexicon as nouns or in the Word Lexicon otherwise 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).

6 MorphoClass System Description

The subsections below correspond to the steps of text processing.

6.1 Unknown Word Tokens and Types Identification

MorphoClass deals with 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)

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. Figure 2 illustrates the decision process, SUB is the tag for substantive NN and EIG - the tag for proper noun PN.

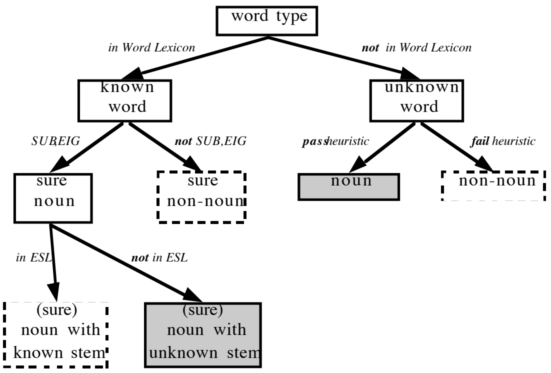

Figure 2. Unknown word/noun/stem identification decision tree

Heuristic

A word *cannot* be a noun iff:

1) $CF = 0$

or

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

where:

t is an empirically chosen, appropriate constant between 0 and 1.

6.2 Generation of all possible stems for unknown nouns

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 into account the umlauts and the *ß* alternations. For each word type all acceptable rule inversions are performed. For example for the word *Lehrerinnen* the following candidate-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 non-empty. 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 correct stemming then all its corresponding inflected word type forms present in the analysed text are grouped together as shown in
Table **5**.

We have 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 of
Table 5. The stem *Haus* is the correct stem for the word *das Haus*, whose morphological class is *n22*. All the wordforms listed there are correct except *Hausse* and *Hausen*. The latter are valid candidates for this stem according to the inflection reversal rules but are incompatible with the correct morphological class *n22*.

MorphoClass does 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 that the stem is correct.*

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 }
Schrei	**3**	**{ Schrei, Schreien, Schreier}**

Table 5. Largest “coverage” stems, ordered by the number of “covered” word types

6.3 Stem coverage refinements

MorphoClass goes through the stems, as generated in Table 3, and for each stem in column one it checkes whether there exists a morphological class that could generate all the wordforms listed in column two. If at least one is found we accept the current coverage; 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 as members of the same paradigm. 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 wordforms it covers taken together, as if they were members of its paradigm. As an example that a candidate-stem can be incorrect consider the wordform *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, there is exactly one correct stem: *Tag*.

How to refine Table 3 rows? An obvious (but not very wise) solution is just to reject the stem which seem to cover “contradicting” wordforms. 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 shown in Table 3. 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 of wordforms - *Spiel*, *Spiele*, *Spielen* and *Spiels* - might be generated by the classes *m1*, *m9* (and *m9a* that has been conflated to *m9*), *n20* and *n25*, while the second one - *Spieler*, *Spielern* - 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 wordforms from *Spiel*? And if we have to choose between the two groups why will we reject the latter one? The most obvious answer is simply because the first group is bigger and thus it is more likely to be correct. If the two groups had the same number of members, we would take the most likely morphological class, which appears more frequently according to the statistics collected from Morphy's lexicon. In the worst case **MorphoClass** would guess two candidates for morphological classification with equivalent likelihood.

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 a 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 out of 13,147 stems. In our opinion, it is even more unlikely that a new unknown word will have more than one morphological class, and additionally is used with two or more of these classes at the same text. We thus always look for only one paradigm for the given the stem. And we always prefer the biggest wordforms set that a morphological class could cover.

Let us consider in more detail the interesting 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 as a frequency of usage of the morphological class.

Table 4a and Table 4b illustrate the refinement algorithm at work. Table 4a lists the top unknown stems found in the NEGRA corpus ordered by the number of covered wordforms (and then alphabetically). It contains quite common wordls like *Ost* and *West*, whose stems were not automatically recognised from the Morphy lexicon and were not included as known in MorphoClass lexicons. Table 4b shows the same list after some refinement.

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, Gemeindehaushal-tes, 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 }**
......		.....
Bildungsurlaube	**3**	**{ Bildungsurlaube, Bildungsurlauben, Bildungsurlauber }**

Table 6a. Unknown stems: ordered by the number of the covered word types

Unknown Stem	#	Words that *Generated* 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 }
Bildungsurlaub	**3**	**{ Bildungsurlaub, Bildungsurlaube, Bildungsurlauben }**
Bildungsurlaube	**3**	**{ Bildungsurlaub, Bildungsurlauben, Bildungsurlauber }**
Bo	**3**	{ Bo, Bose, Boses }
Ost	**3**	**{ Ost, Oster, Ostern }**
West	**3**	**{ West, Weste, Westen }**

Table 7b. Refined stems from table 4a

6.4 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 correct stem for the word types it covers. The morphological analysis exploits both lexicon-based and suffix-based morphology.

6.4.1 Lexicon-based morphology

- *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 we know the morphological class of all the stems in the Stem Lexicon. 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 could generate are already known.

- *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.

6.4.2 Suffix-based 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.

6.4.3 Ending-based morphology

We implemented Mikheev-style rules for ending guessing (Mikheev, 1997). He originally made this for PoS guessing but we applied the same approach for morphological class guessing. We selected a 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.

Table 8 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 four 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 chosen. If more than one of these happened the highest likelihood label has been taken 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 followed by the morphological class or classes it predicts. It is possible that there is more than one class predicted by a single stem (see *Stadtteil*, Table 8) or more than one stems a compound can be split into (see *Gemeindehaushalt*, Table 8). 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.

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, 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 )`

Table 8. Unknown stems with morphological information.

7 Evaluation of the Coverage and Precision of the Rules

MorphoClass system was manually evaluated over four kinds of texts:

Reuters news, a data set of short texts containing 149 different wordforms (word types?), 174 word tokens, 1.43 KB;
Franz Kafka's *Erzählungen*, 3510 wordforms, 13793 word tokens, 85KB;
Goethe's *Die Wahlverwandtschaften*, 10833 wordforms, 79485 word tokens, 517KB
Goethe's *Wilhelm Meisters Lehrjahre*, 17252 wordforms, 194266 wordtokens, 1211 KB.

As we said before, MorphoClass considers some stems as candidate-nouns (normally these candidates include proper nouns, foreign words etc.) and tries to decide which is the inflexional class of the noun. Sometimes assignment is impossible (mostly when only one wordform is met in the text) and then MorphoClass indicates "no info", which means that there is not enough information of how to assign an inflexional class since neither the compound-splitting rule nor the ending-guessing rule were applicable. This is a positive feature of MorphoClass, since it avoids misleading decisions in case of absent information. Table 1 summarises MorphoClass reactions for the four testing data sets. Note that the high percentage of "no info" in the Reuters news may be explained with the numerous foreign names in these texts. We should emphasize that MorphoClass always proposes candidate classes but does not choose one of them. in cases of "no info".

	Stems recognised as compounds	Stems treated by ending-guessing rules	"No info"-stems	Stems recognised as candidate nouns
Reuters	52 (26%)	57 (28%)	91 (46%)	200
Kafka	185 (39%)	190 (40%)	98 (21%)	473
Goethe	551 (32%)	837 (49%)	318 (19%)	1706
W. Meister	896 (32%)	1274 (45%)	668 (23%)	2838

Table 7. MorphoClass reactions according to text types

MorphoClass performs the morphological analysis using both ***compound-splitting*** as well as ***ending-guessing*** rules. These were run in a ***cascade*** manner: the ending-guessing rules were applied *only* if the compound-splitting rules failed. Not surprisingly the compound-splitting rules have coverage of more than 32%, which gives an idea of how often the compound nouns occur on German. Their precision is higher than 92% for all text types. Substantial amount of the ***remaining*** stems, i.e. stems that were not treated by compound-splitting rules - are covered by the ending-guessing rules. Table 1 shows that in case of longer literary texts, ending rules are applied for more than 40% of the stems, in average 45%. Their precision was much lower (see details below).

It should be noted, however, that MorphoClass has no dictionary of named entities and that its ending rules were learnt over the relatively small lexicon of Morphy where the nominalised verbs constitute a considerable part of the dictionary entries. Therefore, we do not pretend that the ending rules applied at present are representative statistics about the possible ending of German nouns. All results given below should be considered as relative, according to the available resources. No doubt a list of named entities and better initial lexicon would influence considerably the results presented here.

A very detailed evaluation of coverage and precision was done using the 85KB text of *Erzählungen* by Franz Kafka. There were 3510 different wordforms found: 862 known nouns, 2155 known non-nouns and *493 unknown nouns*. MorphoClass made the hypothesis that these 493 wordforms had been produced by 473 stems (root forms): there was one stem with 4 word forms, another one with 3 word forms, 15 stems with 2 word forms and the rest — with just one word form. We classified the stems in the following categories:

> SET — A *set* of classes has been assigned. Usually, MorphoClass assigns a single class but in other cases it is a non-empty set of up to 39 classes. About 10% of all stems were in the group of SET;
>
> PART — MorphoClass discovered a *correct* class but *not all* the correct classes. PART has value of 0,6%;
>
> WRONG — MorphoClass assigned a single class and it was *wrong*. About 15% of the stems were WRONGly recognised;
>
> YES — MorphoClass assigned a single class and it was the only correct one. About 60% of the stems were correctly recognised;
>
> SKIP— The stem has been skipped from the current evaluation. We did so for the proper nouns, non-German nouns, non-nouns or in case of incorrect stem. About 10% of all stems were excluded from the evaluation due to these reasons.

We evaluated the System in terms of *precision* and *coverage* in the following way, according to four measures:

precision1 = YES / (YES + WRONG + PART)
precision2 = (YES + (*scaled_PART*)) / (YES + WRONG + PART)
precision3 = (YES + PART) / (YES + WRONG + PART)
coverage = (YES + WRONG + PART) / (YES + WRONG + PART + SET)

The *coverage* shows the proportion of the stems whose morphological class has been found, while the *precision* reveals how correct it was. A scaling is performed according to the proportion of possible classes guessed to the total classes count: if a stem belongs to k (k?2) classes and MorphoClass found one of them (it finds exactly one) *precision1* considers this as a failure (will add 0), *precision2* counts it as a partial success (will add *scaled_PART*=1/k) and *precision3* accepts it as a full success (will add 1).

	RUN 1			RUN 2
	Compound splitting rules	Ending-guessing Rules	Overall (cascade application)	Overall (ending only, with dis-abled com-pounds)
coverage	*43.119266%*	*45.871560%*	*88.990826%*	*76.146789 %*
precision1	93.617021%	56.000000%	74.226804%	66.265060 %
precision2	93.617021%	57.470000%	76.082474%	68.433735 %
precision3	*93.617021%*	*70.000000%*	*81.443299%*	*74.698795 %*

Table 8. MorphoClass system evaluation using Kafka's *Erzaehlungen*. Note that the coverage is higher than in Table 1, since "No info" from Table 1 is split into SET, PART and SKIP

Compound-splitting rules have a very high precision: 93.62% (no partial matching: all the rules considered predicted just one class even when more than one splitting was possible) and coverage of 43.12%. Ending-guessing rules have much lower precision: 56% for *precision1* and 70% for *precision3*. This gives us an overall coverage of 88.99% and precision of 74.23%, 76.08% and 81.44%.

Note that the cascade algorithm is "unfair" since it does not give the ending-guessing rules the opportunity to be applied unless the compound-splitting rules had failed. That is why we made a second run with compound-splitting rules disabled and obtained much higher both coverage (76.15%) and precision (66.27%, 68.43%, 74.70%), see Table 2. ???Do we add the appendix??? Appendix 1 contains all ??? endings learnt over the Morphy lexicon. Note that some elements there are short stems, so ending rules might act as compound splitting. This explains why independent runs of ending-guessing rules (without cascade compound splitting) results in the significant improvement of the performance of the ending rules.

We present below one table for the precision of the ending-guessing rules taken in isolation, according to the inflexion classes.

Class	**Number of wrongly guessed stems in case of X wordforms occurences, where**				**Number of correctly guessed stems in case of X wordforms occurences, where**				**Totals**
	X=1	X=2	X=3	X=4	X=1	X=2	X=3	X=4	
m1	33	1	0	0	39	11	5	0	Wrong: 34 Correct: 55
m2	19	0	0	0	15	1	1	0	Wrong: 19 Correct: 17
m3	4	0	0	0	15	1	3	0	Wrong: 4 Correct: 1
m3a	2	0	0	0	0	0	0	0	Wrong: 2 Correct: 0
m4	23	1	0	0	36	6	0	0	Wrong: 24 Correct: 42
m5	2	0	0	0	1	0	0	0	Wrong: 2 Correct: 1
m6	0	0	0	0	3	0	0	0	Wrong: 0 Correct: 3
m7	3	2	0	0	0	0	0	0	Wrong: 5 Correct: 0
m7a	0	1	0	0	1	0	0	0	Wrong: 1 Correct: 1
m7/ f19/ n26	129	16	2	0	31	7	0	0	Wrong: 147 Correct: 38
m8	7	2	0	0	6	0	0	0	Wrong: 9 Correct: 6
m9	0	0	0	0	3	0	0	0	Wrong: 0 Correct: 3
m10	4	0	0	0	0	0	0	0	Wrong: 4 Correct: 0
m11	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0
f12	6	0	0	0	0	0	0	0	Wrong: 6 Correct: 0
f13	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0
f14	0	0	0	0	6	0	0	0	Wrong: 0 Correct: 6
f14a	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0

f15	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0
f15a	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0
f16	17	1	0	0	77	6	0	0	Wrong: 18 Correct: 83
f17	17	0	0	0	296	30	0	0	Wrong: 17 Correct: 326
f18	5	0	0	0	30	3	0	0	Wrong: 5 Correct: 33
f19	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0
n20	24	1	0	1	15	1	1	0	Wrong: 0 Correct: 0
n20a	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0
n21	2	0	0	0	3	0	0	0	Wrong: 2 Correct: 3
n22	7	1	0	0	1	0	0	0	Wrong: 8 Correct: 1
n23	93	4	0	0	115	3	0	0	Wrong: 97 Correct: 118
n23a	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0
n24	4	0	0	0	1	0	0	0	Wrong: 4 Correct: 1
n25	1	0	0	0	0	0	0	0	Wrong: 1 Correct: 0
n26	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0
n27	0	0	0	0	3	1	0	0	Wrong: 0 Correct: 4
n28	2	0	0	0	0	0	0	0	Wrong: 2 Correct: 0
n28a	1	0	0	0	0	0	0	0	Wrong: 1 Correct: 0
n29	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0
n30	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0
n31	0	0	0	0	0	0	0	0	Wrong: 0 Correct: 0

Table 9. Distribution of 1154 wrongly and correctly guessed stems in “Wilhelm Meisters Lehrjahre” (120 cases of guessing no nouns, proper names and non-German words are skipped)

Note that ending-guessing rules are applied (1) to unknown nouns, i.e. nouns outside the lexicon and (2) after compound-splitting rules. So, Table 9 is not a representative statistics about German nouns and their inflexional classes

8 Future work and conclusion

Two new modules were elaborated recently, for integration in **MorphoClass**, but unfortunately we can currently present partial results only instead of exhaustive evaluation.

7.1 Further development: Integration of 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 multi-graph since each stem could be generated by more than one word form and each word form 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 a more reliable morphological information attached is preferred. This means that 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 a stem that is covered by an ending guessing rule. (The known stems are simply rejected, see above).
4) All other being equal, a longer stem is preferred.

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)

Table 10. *Selected* unknown stems together with the morphological information available till now. (NEGRA corpus)

7.2 Further development: Context processing

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. Thå article *das* predicts that 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. 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
- pronouns: possessive, demonstrative, indefinite

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 the 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.

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 check the possibility that the morphological class *m3a* is acceptable. We consider 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.

After the word types have been covered by stems we build vectors for each separate word type. The vector has 24 ($3 \times 2 \times 4$) 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 co-ordinates 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 a vector (all vector co-ordinates will be 0).

7.3 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 with almost no 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.

7.4 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. This 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.

9 References - one more needed (at least)

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)

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, IOS Press, Vol. 35 in the series "Frontiers in AI and Applications", pp. 218 - 227.

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, Vol. 1115, pp. 115 - 129.

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.

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/cutting92 practical.html)

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

DeJean H. (1998) *Morphemes as necessary concepts for structures: Discovery from untagged corpora.* University of Caen-Basse Normandie. 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.

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)

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.

v. Hahn W. (2002) *HyperLAT, the DBR-MAT Lexicon Acquisition tool. The Manual*
(http://nats-www.informatik.uni-hamburg.de/~vhahn/vHahnLit.html)

v. Hahn W. (1999) *Metamodelling of Lexical Acquisition Tools. In: Proceedings of Eurolan'99. Iasi 1999. pp. 197 -201*

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.

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

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)

Kazakov D. (1997) *Unsupervised Learning of Naïve 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. (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.

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

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)

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)

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)

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)

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.

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)

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)

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.

Van den Bosch, A. and W. Daelemans. (1999) *Memory-based morphological analysis.* Proc. of the 37th An. Meeting of the ACL, University of Maryland, pp. 285-292.

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.

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)

10 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(R)-MAT project
http://nats-www.informatik.uni-hamburg.de/~dbrmat/ ; http://www.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 Appendix 1

Here follow the Mikheev-style ending-guessing rules. He originally applies them for guessing POS for unknown words but we use a similar approach for our morphological classes.

Each line contains a single ending-guessing rule. The line starts with an ending (1 to 6 letters), then follows its confidence level (acoording to the Mikheev formula), then follows a list of morpshological classe each followed by the frequency it occured with the ending considered. If more than one morphological classes are listed the most frequent one is the one the predicted by the rule. the others are just a noise.

Ending-guessing rules with conbfidence level of at least 90% and a frequency of at least 11 are considered. the length allowed is set to up to 7 letters.

In fact the confidence level is important just during the ending-guessing rules generation phase and is unnecessary later: The rules are applied in cascade manner. We checek whether there is an eneding-guessing rules matching the last 7 letters. If so, we assign the class it predicts. Otherwise, we check for a rule matching the last 6 letters, then the last 5 etc.

---

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 5
		f17 3822
len	0.998139	m4 510
ndung	0.998131	f17 455
ion	0.998114	m1 1
		f17 1207
egung	0.998111	f17 450
tling	0.998032	m1 432
uß	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

öhl	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
loß	0.997005	n22 317
spiel	0.996989	n20 282
fnung	0.996989	f17 282
anken	0.996989	m4 282
hrung	0.996957	f17 279
eger	0.996931	m4 289
luß	0.996907	m2 307
piel	0.996877	n20 284
mung	0.996855	f17 282
riff	0.996833	m1 280
ner	0.996824	m4 1101
		n23 2
age	0.99676	f16 293
iel	0.996658	n20 284
derung	0.996599	f17 242
zung	0.99651	f17 254
rer	0.99625	m4 253
altung	0.996244	f17 219
ute	0.996189	f16 249
allen	0.996178	m4 222
ruch	0.996164	m2 231
ltung	0.996144	f17 220
mmung	0.996144	f17 220
chäft	0.996126	n20 219
such	0.996113	m1 228
enstand	0.996062	m2 204
llen	0.996044	m4 224
hnung	0.996036	f17 214
rin	0.996014	f18 238
nnung	0.99598	f17 211
ulein	0.99598	n24 211
cher	0.995973	m4 220
nstand	0.995969	m2 204
wester	0.995969	f16 204
häft	0.995954	n20 219
samkeit	0.995938	f17 198
ig	0.995936	m1 266
ahme	0.995898	f16 216
erin	0.995879	f18 215
amkeit	0.995842	f17 198
rag	0.995803	m2 226
gesicht	0.995768	n21 190
sung	0.995762	f17 209
icksal	0.995713	n20 192
mkeit	0.995712	f17 198
findung	0.9957	f17 187
äft	0.99567	n20 219
esicht	0.995668	n21 190
ilie	0.995658	f16 204
nahme	0.995646	f16 195
hme	0.99561	f16 216
cksal	0.995578	n20 192
aße	0.995527	f16 212
cke	0.995485	f16 210
dlung	0.995485	f17 188
satz	0.995476	m2 196
lie	0.995442	f16 208
tschaft	0.995433	f17 176
lick	0.995429	m1 194
ksal	0.995382	n20 192
sion	0.995382	f17 192
undin	0.995285	f18 180
ngung	0.995178	f17 176
aft	0.995162	f14 5
		f17 1439
chied	0.995123	m1 174
ndin	0.995075	f18 180
chlag	0.995067	m2 172
merz	0.994964	m9 176
ierung	0.994954	f17 163
tzung	0.99495	f17 168
unft	0.994935	f14 175
warze	0.99492	f16 167
bindung	0.994916	f17 158
haltung	0.994916	f17 158
hied	0.994907	m1 174
eler	0.994907	m4 174
pruch	0.994889	m2 166
zeß	0.99487	m1 185
din	0.99487	f18 185
rkung	0.994859	f17 165
tler	0.994848	m4 172
hlag	0.994848	m2 172
ker	0.994786	m4 182
tät	0.994728	f17 180
arze	0.994694	f16 167
trag	0.994662	m2 166
nft	0.994578	f14 175
chluß	0.994458	m2 153
eiter	0.994458	m4 153
gnügen	0.994408	n23 147
stag	0.994394	m1 158
eß	0.994339	m1 191
iene	0.994322	f16 156
iter	0.994286	m4 155
ührung	0.994253	f17 143
nügen	0.994233	n23 147
itz	0.994217	m1 164
schlag	0.994213	m2 142
hluß	0.994211	m2 153
nnerung	0.994184	f17 138
achtung	0.994099	f17 136
erheit	0.994046	f17 138
nerung	0.994046	f17 138
itär	0.994017	f17 148
spieler	0.994011	m4 134

tigkeit	0.994011	f17 134	zimmer	0.992471	n23 109
digung	0.994003	f17 137	urf	0.992425	m2 125
ügen	0.993977	n23 147	itekt	0.992374	m8 111
lion	0.993977	f17 147	stler	0.992374	m4 111
herheit	0.993875	f17 131	gene	0.992309	m7 115
erkung	0.993869	f17 134	eitung	0.99226	f17 106
pieler	0.993869	m4 134	schluß	0.99226	m2 106
rakter	0.993824	m1 133	ule	0.99224	f16 122
rechen	0.993824	n23 133	hlung	0.992235	f17 109
ehung	0.993815	f17 137	stück	0.992235	n20 109
idung	0.99377	f17 136	pf	0.992225	m1 1
dschaft	0.993732	f17 128			m2 321
fernung	0.993683	f17 127	heinung	0.992221	f17 103
iehung	0.993682	f17 130	eilung	0.992187	f17 105
ieler	0.993677	m4 134	ergang	0.992112	m2 104
tigung	0.993633	f17 129	einung	0.992112	f17 104
akter	0.99363	m1 133	eresse	0.992036	n23 103
echen	0.99363	n23 133	tekt	0.992034	m8 111
innung	0.993584	f17 128	halt	0.992034	m1 111
ernung	0.993534	f17 127	gramm	0.992017	n20 106
cheid	0.993484	m1 130	itung	0.992017	f17 106
gabe	0.993445	f16 135	kunft	0.992017	f14 106
utung	0.993434	f17 129	eb	0.992005	m1 135
rnung	0.993383	f17 128	rschaft	0.991989	f17 100
rgang	0.993383	m2 128	andlung	0.991989	f17 100
hricht	0.993379	f17 124	zug	0.991974	m2 118
ifel	0.993347	m4 133	tur	0.991974	f17 118
kter	0.993347	m1 133	äude	0.991963	n23 110
mnis	0.993297	n27 132	fung	0.991963	f17 110
enz	0.993279	f17 141	ilung	0.991942	f17 105
mittag	0.993216	m1 121	tück	0.991889	n20 109
rter	0.993194	m4 130	inung	0.991865	f17 104
heid	0.993194	m1 130	ite	0.991837	f16 116
ohner	0.993171	m4 124	timmung	0.991826	f17 98
ition	0.993171	f17 124	lage	0.991815	f16 108
eimnis	0.993159	n27 120	nitt	0.991815	m1 108
kfurter	0.993144	m4 117	ndlung	0.9918	f17 100
ichtung	0.993144	f17 117	eidung	0.991716	f17 99
inn	0.993084	m1 137	tten	0.991662	m4 106
nze	0.993084	f16 137	ramm	0.991662	n20 106
ittag	0.993003	m1 121	etzung	0.991632	f17 98
zer	0.992983	m4 135	immung	0.991632	f17 98
furter	0.992982	m4 117	iner	0.991583	m4 105
hner	0.99298	m4 126	etz	0.991548	n20 112
rift	0.99298	f17 126	heimer	0.991546	m4 97
nderung	0.992965	f17 114	lige	0.991503	f16 104
imnis	0.992945	n27 120	denheit	0.991388	f17 93
wurf	0.992924	m2 125	rf	0.99137	m2 125
führung	0.992904	f17 113	ück	0.991317	n20 109
angene	0.992861	m7 115	suchung	0.991296	f17 92
hang	0.99281	m2 123	eimer	0.991281	m4 97
urter	0.992762	m4 117	uchung	0.991279	f17 94
ttag	0.992692	m1 121	eutung	0.991279	f17 94
ecke	0.992692	f16 121	fall	0.991253	m2 101
ngene	0.992637	m7 115	zeit	0.991253	f17 101
auer	0.992631	m4 120	ohnheit	0.991201	f17 91
chnung	0.992606	f17 111	schung	0.991184	f17 93
hitekt	0.992606	m8 111	ef	0.991132	m1 426
ulter	0.992573	f16 114			n24 2
ahrung	0.992471	f17 109	iheit	0.991099	f17 95

engung	0.991089	f17 92
tei	0.991073	f17 106
ament	0.991005	n20 94
hnheit	0.990992	f17 91
weg	0.990989	m1 105
olg	0.990989	m1 105
rtung	0.990907	f17 93
rengung	0.990902	f17 88
imer	0.990893	m4 97
enteil	0.990792	n20 89
hauer	0.990709	m4 91
rschied	0.990693	m1 86
rnehmen	0.990693	n23 86
hstück	0.990686	n20 88
steck	0.990607	n20 90
ntag	0.990605	m1 94
bung	0.990605	f17 94
hmittag	0.990584	m1 85
prechen	0.990584	n23 85
ife	0.990542	f16 100
bot	0.990542	n20 100
ktion	0.990503	f17 89
nteil	0.990503	n20 89
teck	0.990502	n20 93
ligkeit	0.990473	f17 84
schied	0.990472	m1 86
nehmen	0.990472	n23 86
ener	0.9904	m4 92
gling	0.990284	m1 87
ehmen	0.990172	n23 86
tritt	0.990172	m1 86
die	0.990149	f16 96
lärung	0.990129	f17 83
chauer	0.990129	m4 83
iergang	0.990122	m2 81
teilung	0.990122	f17 81
sa	0.990107	m6 109
ehen	0.99008	n23 89
iet	0.990046	n20 95
uck	0.990031	m1 349
		m2 2
iker	0.989965	m4 88
ehl	0.989941	m1 94
izier	0.989941	m1 84
ingung	0.989888	f17 81
ing	0.989884	m1 685
		m6 1
		f15 1
		n20 1
		n24 2
ruck	0.989882	m1 330
		m2 2
ium	0.989831	n28 93
ssung	0.989818	f17 83
ärung	0.989818	f17 83
spruch	0.989763	m2 80
assung	0.989763	f17 80
meister	0.989744	m4 78
hmen	0.989734	n23 86
nger	0.989734	m4 86

lg	0.989733	m1 105
önheit	0.989632	f17 79
eden	0.989615	m4 85
fluß	0.989615	m2 85
auf	0.98961	m2 91
glied	0.98957	n21 81
eister	0.989501	m4 78
liarde	0.989501	f16 78
zier	0.989492	m1 84
rie	0.989379	f16 89
eugung	0.989366	f17 77
chrift	0.989366	f17 77
ugung	0.989306	f17 79
aden	0.989236	m5 82
nkheit	0.989227	f17 76
lnahme	0.989227	f16 76
zeugung	0.989194	f17 74
enthalt	0.989194	m1 74
ignis	0.98917	n27 78
iarde	0.98917	f16 78
uung	0.989105	f17 81
lied	0.989105	n21 81
hrift	0.989031	f17 77
mögen	0.989031	n23 77
nthalt	0.988937	m1 74
kheit	0.988889	f17 76
ldung	0.988889	f17 76
einde	0.988889	f16 76
herung	0.988787	f17 73
ration	0.988787	f17 73
thalt	0.988589	m1 74
mlung	0.988589	f17 74
dheit	0.988589	f17 74
ögen	0.988542	n23 77
dent	0.988542	m8 77
ung	0.988524	m2 109
		f17 10009
rkeit	0.988435	f17 73
inde	0.988393	f16 76
hs	0.988288	m2 92
atten	0.988273	m4 72
sal	0.988224	f12 1
		n20 192
nkt	0.98819	m1 80
ürfnis	0.98814	n27 69
iener	0.98811	m4 71
zicht	0.98811	m1 71
setzung	0.988074	f17 67
tie	0.988039	f16 79
rdnung	0.987968	f17 68
ählung	0.987968	f17 68
olver	0.987942	m4 70
unde	0.987919	f16 73
hren	0.987919	n23 73
rze	0.987848	f16 186
		n23 1
rfnis	0.987767	n27 69
äche	0.98775	f16 72
fte	0.987732	f16 77
menhang	0.98771	m2 65

dnung	0.987589	f17 68
enhang	0.987417	m2 65
dernis	0.987417	n27 65
liner	0.987406	m4 67
tin	0.987405	f18 75
onin	0.987404	f18 70
benheit	0.987322	f17 63
fnis	0.98722	n27 69
eg	0.98715	m1 293
		n24 1
		n31 1
nhang	0.987021	m2 65
tive	0.986844	f16 67
punkt	0.986821	m1 64
ise	0.9867	f16 71
ftigung	0.986695	f17 60
nzessin	0.986695	f18 60
lität	0.986611	f17 63
änger	0.986611	m4 63
legung	0.986598	f17 61
elheit	0.986598	f17 61
eis	0.986562	m1 168
		n20 1
nin	0.986512	f18 70
hbar	0.986441	m7 65
scher	0.986398	m4 62
fahren	0.986378	n23 60
zessin	0.986378	f18 60
ive	0.986315	f16 69
rigkeit	0.98624	f17 58
unkt	0.986232	m1 64
enken	0.986175	n23 61
lheit	0.986175	f17 61
äusch	0.986175	n20 61
per	0.986155	m4 163
		f16 1
gin	0.986117	f18 68
igin	0.986013	f18 63
rund	0.986013	m2 63
essin	0.985948	f18 60
vater	0.985948	m5 60
ahren	0.985948	n23 60
uhe	0.985912	f16 67
artung	0.985912	f17 58
chmack	0.985912	m3 58
ersehen	0.985753	n23 56
nzimmer	0.985753	n23 56
ommen	0.98571	n23 59
ügung	0.98571	f17 59
terung	0.985665	f17 57
länder	0.985665	m4 57
ade	0.985634	m7 6
		f16 635
		n23 1
tzer	0.985557	m4 61
enk	0.98548	n20 65
hmack	0.985467	m3 58
ikant	0.985467	m8 58
tisch	0.985467	m1 58
rsehen	0.985413	n23 56
nteuer	0.985413	n23 56
iger	0.98532	m4 60
aber	0.98532	m4 60
hof	0.985256	m2 64
mpf	0.985256	m2 64
ächtnis	0.98523	n27 54
stadt	0.985213	f14 57
gleich	0.985149	m1 55
uation	0.985149	f17 55
mmen	0.985071	n23 59
tner	0.985071	m4 59
heidung	0.984953	f17 53
ckung	0.984953	f17 56
sehen	0.984953	n23 56
stung	0.984953	f17 56
teuer	0.984953	n23 56
bruch	0.984953	m2 56
nehmer	0.984878	m4 54
chtnis	0.984878	n27 54
kant	0.984817	m8 58
mack	0.984817	m3 58
leich	0.98468	m1 55
lerin	0.98468	f18 55
blick	0.98468	m1 55
chlecht	0.984668	n21 52
ksicht	0.984594	f17 53
ndheit	0.984594	f17 53
echung	0.984594	f17 53
tadt	0.984551	f14 57
erei	0.984551	f17 57
itän	0.984551	m1 57
htnis	0.9844	n27 54
haber	0.9844	m4 54
ehmer	0.9844	m4 54
tchen	0.9844	n23 54
rderung	0.984369	f17 51
hundert	0.984369	n20 51
undheit	0.984369	f17 51
fügung	0.984303	f17 52
sation	0.984303	f17 52
bhaber	0.984303	m4 52
hlecht	0.984303	n21 52
leiter	0.984303	m4 52
ebnis	0.984107	n27 53
wäche	0.984107	f16 53
chenk	0.984107	n20 53
tiker	0.984107	m4 53
erstand	0.984059	m2 50
aubnis	0.983996	f13 51
nsucht	0.983996	f14 51
sition	0.983996	f17 51
undert	0.983996	n20 51
hsel	0.983994	m4 55
tier	0.983994	n20 55
aum	0.98374	m2 58
reitung	0.983739	f17 49
osition	0.983739	f17 49
hmer	0.983702	m4 54
zeug	0.983702	n20 54
rstand	0.983678	m2 50

lüssel	0.983678	m4 50
chnitt	0.983678	m1 50
ubnis	0.98349	f13 51
ssion	0.98349	f17 51
sucht	0.98349	f14 51
ndert	0.98349	n20 51
adt	0.983454	f14 57
tän	0.983454	m1 57
hse	0.983448	f16 136
		n23 1
rspruch	0.983403	m2 48
pich	0.983395	m1 53
henk	0.983395	n20 53
erie	0.983395	f16 53
kommen	0.983351	n23 49
eckung	0.983351	f17 49
ildung	0.983351	f17 49
regung	0.983351	f17 49
scheid	0.983351	m1 49
of	0.983194	m2 64
sende	0.983162	f16 50
hnitt	0.983162	m1 50
haben	0.983162	n23 50
blem	0.983081	n20 52
urtstag	0.983052	m1 47
schrift	0.983052	f17 47
digkeit	0.983052	f17 47
it	0.983051	m1 38
		m8 7
		f17 3822
		n20 6
		n24 5
		n25 3
rrung	0.982824	f17 49
dert	0.982751	n20 51
ette	0.982751	f16 51
wirrung	0.982691	f17 46
ichnung	0.982691	f17 46
fassung	0.982691	f17 46
rtstag	0.982647	m1 47
eug	0.982545	n20 54
hzeit	0.982469	f17 48
aben	0.982407	n23 50
schnitt	0.98231	m1 45
itiker	0.982277	m4 46
rikant	0.982277	m8 46
irrung	0.982277	f17 46
rre	0.982216	f16 53
nie	0.982216	f16 53
ucher	0.982098	m4 47
tstag	0.982098	m1 47
üler	0.982054	m4 49
anze	0.982054	f16 49
ulde	0.982054	f16 49
deckung	0.981912	f17 44
reibung	0.981912	f17 44
ßvater	0.981887	m5 45
ntasie	0.981887	f16 45
acher	0.981716	m4 46
ative	0.981716	f16 46

tasie	0.981716	f16 46
arre	0.981682	f16 48
eber	0.981682	m4 48
dler	0.981682	m4 48
ebot	0.981682	n20 48
werk	0.981682	n20 48
eur	0.981525	m1 51
rbeiter	0.981496	m4 43
rechung	0.981496	f17 43
cherung	0.981496	f17 43
eibung	0.981479	f17 44
re	0.981474	f16 268
		n24 3
tnant	0.981313	m6 45
erad	0.981295	m8 47
än	0.981136	m1 57
rde	0.981125	m7 1
		f16 119
bildung	0.98106	f17 42
brechen	0.98106	n23 42
icklung	0.98106	f17 42
trument	0.98106	n20 42
ndliche	0.98106	m7 42
beiter	0.981053	m4 43
asie	0.980896	f16 46
zert	0.980896	n20 46
ibung	0.980892	f17 44
liche	0.980892	m7 44
ühung	0.980892	f17 44
tik	0.980779	f17 49
ieb	0.980779	m1 49
rument	0.980606	n20 42
dliche	0.980606	m7 42
ensatz	0.980606	m2 42
cklung	0.980606	f17 42
rlegung	0.980603	f17 41
isation	0.980603	f17 41
nant	0.980474	m6 45
itzer	0.980452	m4 43
atung	0.980452	f17 43
ftung	0.980452	f17 43
issen	0.980242	m4 3
		n23 220
sident	0.980139	m8 41
alität	0.980139	f17 41
chheit	0.980139	f17 41
wohner	0.980139	m4 41
ldigung	0.980125	f17 40
aschung	0.980125	f17 40
erricht	0.980125	m1 40
nheimer	0.980125	m4 40
uz	0.980099	n20 54
onie	0.980034	f16 44
nese	0.980034	m7 44
rieb	0.980034	m1 44
jekt	0.980034	n20 44
nsatz	0.979991	m2 42
izont	0.979991	m1 42
klung	0.979991	f17 42
ssen	0.979978	m4 3

		n23 222	üsch	0.977502	n20 39
egnung	0.979648	f17 40	raum	0.977502	m2 39
gänger	0.979648	m4 40	irge	0.977502	n23 39
rricht	0.979648	m1 40	ißheit	0.977415	f17 36
tation	0.979648	f17 40	diener	0.977415	m4 36
itel	0.979636	m4 9	mation	0.977415	f17 36
		n23 539	nigung	0.977415	f17 36
nerstag	0.979622	m1 39	schheit	0.977325	f17 35
lauf	0.979574	m2 43	ernacht	0.977325	f14 35
itag	0.979574	m1 43	terkeit	0.977325	f17 35
ober	0.979574	m4 43	schritt	0.977325	m1 35
sie	0.979537	f16 46	chick	0.977323	n20 37
prung	0.979509	m2 41	ndnis	0.977323	n27 37
hheit	0.979509	f17 41	robe	0.976917	f16 38
ident	0.979509	m8 41	beit	0.976917	f17 38
gnung	0.979509	f17 41	umpf	0.976917	m2 38
tament	0.979133	n20 39	rist	0.976917	m8 38
erstag	0.979133	m1 39	eine	0.976917	m7 38
ibtisch	0.979093	m1 38	liener	0.97678	m4 35
nigkeit	0.979093	f17 38	erkeit	0.97678	f17 35
chafter	0.979093	m4 38	rnacht	0.97678	f14 35
barkeit	0.979093	f17 38	üllung	0.97678	f17 35
ödie	0.979092	f16 42	chritt	0.97678	m1 35
zont	0.979092	m1 42	ßheit	0.976697	f17 36
tang	0.979092	m6 42	ution	0.976697	f17 36
rede	0.979092	f16 42	ndent	0.976697	m8 36
ere	0.979085	f16 45	geber	0.976697	m4 36
lacht	0.979002	f17 40	echnung	0.97667	f17 34
terin	0.979002	f18 40	ergrund	0.97667	m2 34
nkung	0.979002	f17 40	rug	0.976493	m2 40
ähr	0.978613	f17 44	afe	0.976493	f16 40
hafter	0.978592	m4 38	eife	0.976301	f16 37
arbeit	0.978592	f17 38	hick	0.976301	n20 37
endung	0.978592	f17 38	rd	0.976261	m1 10
btisch	0.978592	m1 38			m6 10
arkeit	0.978592	f17 38			n20 998
älde	0.978587	n23 41	weizer	0.976109	m4 34
pfen	0.978587	m4 41	rgrund	0.976109	m2 34
kelheit	0.978537	f17 37	sagier	0.976109	m1 34
tändnis	0.978537	n27 37	litten	0.976109	m4 34
hling	0.978471	m1 39	reuung	0.976109	f17 34
rstag	0.978471	m1 39	dition	0.976109	f17 34
tem	0.978119	n20 43	erlage	0.976109	f16 34
lp	0.978081	m1 49	hritt	0.976041	m1 35
rger	0.978058	m4 40	örung	0.976041	f17 35
gerung	0.978021	f17 37	grund	0.976041	m2 35
ändnis	0.978021	n27 37	lkeit	0.976041	f17 35
rdiener	0.977945	m4 36	lkerung	0.975971	f17 33
leitung	0.977945	f17 36	ammlung	0.975971	f17 33
taltung	0.977945	f17 36	tendent	0.975971	m8 33
echer	0.977912	m4 38	lassung	0.975971	f17 33
sache	0.977912	f16 38	tenz	0.975646	f17 36
rbeit	0.977912	f17 38	bild	0.975646	n21 36
nacht	0.977912	f14 38	mmlung	0.975393	f17 33
gerin	0.977912	f18 38	kerung	0.975393	f17 33
after	0.977912	m4 38	endent	0.975393	m8 33
bst	0.977602	m1 42	prache	0.975393	f16 33
ont	0.977602	m1 42	bacher	0.975393	m4 33
pfer	0.977502	m4 39	mutung	0.975393	f17 33
erer	0.977502	m4 39	hmung	0.975349	f17 34

rlage	0.975349	f16 34
euung	0.975349	f17 34
zchen	0.975349	n23 34
itten	0.975349	m4 34
agier	0.975349	m1 34
eizer	0.975349	m4 34
erb	0.975269	m1 38
obe	0.975269	f16 38
ept	0.975269	n20 38
inigung	0.975229	f17 32
ller	0.974961	m4 35
weis	0.974961	m1 35
nhof	0.974961	m2 35
elkeit	0.974633	f17 32
hstabe	0.974633	m7 32
penst	0.974609	n21 33
bling	0.974609	m1 33
rache	0.974609	f16 33
idigung	0.974447	f17 31
lnehmer	0.974447	m4 31
gier	0.974236	m1 34
eral	0.974236	m2 34
izer	0.974236	m4 34
odie	0.974236	f16 34
vier	0.974236	n20 34
lle	0.974042	m7 22
		f16 959
one	0.973906	f16 36
sor	0.973906	m9 36
opf	0.973906	m2 36
achten	0.973831	n23 31
eigung	0.973831	f17 31
derobe	0.973831	f16 31
liebe	0.973824	f16 32
stein	0.973824	m1 32
rteil	0.973824	n20 32
stabe	0.973824	m7 32
pfung	0.973824	f17 32
eid	0.973621	m1 130
		n21 2
treuung	0.973608	f17 30
sternis	0.973608	f13 30
fang	0.973462	m2 33
efel	0.973462	m4 33
ist	0.973345	m3a 3
		m8 172
arde	0.973121	m7 1
		f16 79
zent	0.973058	m8 3
		n20 164
gehen	0.972996	n23 31
estät	0.972996	f17 31
chtum	0.972996	m3 31
chten	0.972996	n23 31
stenz	0.972996	f17 31
erobe	0.972996	f16 31
essor	0.972996	m9 31
ndsatz	0.972971	m2 30
treter	0.972971	m4 30
ternis	0.972971	f13 30
tgeber	0.972971	m4 30
edigung	0.972713	f17 29
achsene	0.972713	m7 29
folgung	0.972713	f17 29
okratie	0.972713	f16 29
eige	0.972641	f16 32
ütze	0.972641	m7 32
tabe	0.972641	m7 32
dsatz	0.972108	m2 30
reter	0.972108	m4 30
kzeug	0.972108	n20 30
urteil	0.972054	n20 29
chsene	0.972054	m7 29
lament	0.972054	n20 29
olgung	0.972054	f17 29
istrat	0.972054	m1 29
kratie	0.972054	f16 29
eichen	0.972054	n23 29
gefühl	0.972054	n20 29
rb	0.971786	m1 38
ssor	0.971775	m9 31
stät	0.971775	f17 31
hten	0.971775	n23 31
rium	0.971775	n28 31
eibe	0.971775	f16 31
htum	0.971775	m3 31
haftung	0.971755	f17 28
ptstadt	0.971755	f14 28
le	0.97166	m7 31
		m10 6
		f16 2405
		n23 16
		n24 10
nik	0.971562	f17 33
nne	0.971562	f16 33
lgung	0.971161	f17 29
onung	0.971161	f17 29
hsene	0.971161	m7 29
ichen	0.971161	n23 29
ratie	0.971161	f16 29
efühl	0.971161	n20 29
strat	0.971161	m1 29
monie	0.971161	f16 29
leier	0.971161	m4 29
aftung	0.971073	f17 28
ennung	0.971073	f17 28
nation	0.971073	f17 28
tstadt	0.971073	f14 28
ladung	0.971073	f17 28
nist	0.970846	m8 30
ande	0.970846	f16 30
oner	0.970846	m4 30
eier	0.970846	m4 30
atur	0.970846	f17 30
urt	0.97068	f17 32
gie	0.97068	f16 32
nge	0.970308	m7 1
		f16 75
adung	0.970147	f17 28
griff	0.970147	m1 28

fessor	0.970014	m9 27
sprung	0.970014	m2 27
schule	0.970014	f16 27
tole	0.969854	f16 29
frau	0.969854	f17 29
renz	0.969854	f17 29
lade	0.969854	f16 29
zept	0.969854	n20 29
fühl	0.969854	n20 29
atie	0.969854	f16 29
ibe	0.969752	f16 31
une	0.969752	f16 31
lver	0.969748	m4 70
		n23 1
wortung	0.96962	f17 26
olution	0.96962	f17 26
tützung	0.96962	f17 26
rholung	0.96962	f17 26
sequenz	0.96962	f17 26
fschaft	0.96962	f17 26
nehmung	0.96962	f17 26
andte	0.969381	m7 67
		f16 1
quenz	0.969054	f17 27
nzose	0.969054	m7 27
okrat	0.969054	m8 27
mheit	0.969054	f17 27
nstag	0.969054	m1 27
chule	0.969054	f16 27
resse	0.969015	f16 2
		n23 103
equenz	0.968885	f17 26
ützung	0.968885	f17 26
lution	0.968885	f17 26
holung	0.968885	f17 26
ligung	0.968885	f17 26
folger	0.968885	m4 26
ortung	0.968885	f17 26
terium	0.968885	n28 26
ckwerk	0.968885	n20 26
ehmung	0.968885	f17 26
hler	0.968794	m4 28
urke	0.968794	m7 28
tzen	0.968794	m4 28
kauf	0.968794	m2 28
fahr	0.968794	m8 28
rung	0.968454	m2 41
		f17 1384
ndte	0.968428	m7 67
		f16 1
äherung	0.968426	f17 25
emokrat	0.968426	m8 25
anke	0.968064	m7 340
		f16 9
ember	0.967887	m4 26
olung	0.967887	f17 26
diger	0.967887	m4 26
erium	0.967887	n28 26
hilfe	0.967887	f16 26
olger	0.967887	m4 26

nchen	0.967887	n23 26
ltier	0.967887	n20 26
kchen	0.967887	n23 26
kwerk	0.967887	n20 26
irn	0.96769	n20 29
ruf	0.96769	m1 29
rau	0.96769	f17 29
nda	0.96769	f15 29
mokrat	0.96766	m8 25
sage	0.967649	f16 27
zose	0.967649	m7 27
örer	0.967649	m4 27
uenz	0.967649	f17 27
chuß	0.967649	m2 27
hule	0.967649	f16 27
elte	0.967649	f16 27
jahr	0.967649	n20 27
krat	0.967649	m8 27
derlage	0.967135	f16 24
zelheit	0.967135	f17 24
eration	0.967135	f17 24
sierung	0.967135	f17 24
ispruch	0.967135	m2 24
ränkung	0.967135	f17 24
förster	0.967135	m4 24
rmation	0.967135	f17 24
ppe	0.966984	m7 10
		f16 369
dte	0.96685	m7 67
		f16 1
chine	0.966622	f16 25
ultat	0.966622	n20 25
inett	0.966622	n20 25
ahlin	0.966622	f18 25
nce	0.966552	f16 28
toß	0.966552	m2 28
wur	0.966552	m2 28
rke	0.966552	m7 28
ophe	0.966428	f16 26
nett	0.966428	n20 26
omme	0.966428	m7 26
lger	0.966428	m4 26
ider	0.966428	m4 26
hnis	0.966428	n27 26
mber	0.966428	m4 26
tzchen	0.966337	n23 24
ichnis	0.966337	n27 24
änkung	0.966337	f17 24
achter	0.966337	m4 24
eimrat	0.966337	m2 24
örster	0.966337	m4 24
uktion	0.966337	f17 24
ivität	0.966337	f17 24
merung	0.966337	f17 24
sigkeit	0.965736	f17 23
fehlung	0.965736	f17 23
uhigung	0.965736	f17 23
kussion	0.965736	f17 23
narbeit	0.965736	f17 23
wandte	0.96552	m7 58

		f16 1
sur	0.965322	f17 27
ode	0.965322	f16 27
huß	0.965322	m2 27
rster	0.965255	m4 24
flikt	0.965255	m1 24
vität	0.965255	f17 24
chnis	0.965255	n27 24
imrat	0.965255	m2 24
shalt	0.965255	m1 24
hine	0.965104	f16 25
isse	0.965104	f16 25
örde	0.965104	f16 25
umph	0.965104	m1 25
mann	0.965104	m3 25
elei	0.965104	f17 25
hlin	0.965104	f18 25
ehrung	0.964903	f17 23
ussion	0.964903	f17 23
higung	0.964903	f17 23
lation	0.964903	f17 23
eisung	0.964903	f17 23
ehlung	0.964903	f17 23
pe	0.964465	m7 21
		f16 689
ue	0.964333	f16 30
sterium	0.964207	n28 22
uskript	0.964207	n20 22
ndpunkt	0.964207	m1 22
hschule	0.964207	f16 22
eiligte	0.964207	m7 22
phe	0.964012	f16 26
her	0.963973	m4 278
		f16 8
inder	0.963774	m4 23
inger	0.963774	m4 23
isung	0.963774	f17 23
likt	0.963673	m1 24
rrer	0.963673	m4 24
tage	0.963673	f16 24
list	0.963673	m8 24
name	0.963673	m7 24
anda	0.963673	f15 24
rage	0.963673	f16 24
mrat	0.963673	m2 24
skript	0.963335	n20 22
eheuer	0.963335	n23 22
dpunkt	0.963335	m1 22
elerin	0.963335	f18 22
burger	0.963335	m4 22
iligte	0.963335	m7 22
teller	0.963335	m4 22
da	0.963115	f15 29
ument	0.962674	m8 2
		n20 85
mph	0.962589	m1 25
chtling	0.962538	m1 21
steller	0.962538	m4 21
sprache	0.962538	f16 21
pliment	0.962538	n20 21

mte	0.962485	m7 59
		f16 1
ube	0.962448	m7 10
		m7a 628
		f16 11
fel	0.962433	m4 351
		f16 11
urger	0.962153	m4 22
ligte	0.962153	m7 22
theit	0.962153	f17 22
eller	0.962153	m4 22
riebe	0.962153	n23 22
kript	0.962153	n20 22
heuer	0.962153	n23 22
lchen	0.962153	n23 22
derer	0.962153	m4 22
othek	0.962153	f17 22
hrer	0.962122	m4 23
dukt	0.962122	n20 23
wede	0.962122	m7 23
nzip	0.962122	n31 23
iere	0.962122	f16 23
takt	0.962122	m1 23
wahl	0.962122	f17 23
digt	0.962122	f17 23
rank	0.962122	m2 23
ßung	0.962122	f17 23
ehnung	0.961624	f17 21
liment	0.961624	n20 21
neider	0.961624	m4 21
ülerin	0.961624	f18 21
htling	0.961624	m1 21
ust	0.961578	m1 118
		f14 3
chen	0.96117	m4 42
		n23 1142
auß	0.961051	m2 24
hie	0.961051	f16 24
lecht	0.96084	n20 1
		n21 52
ätigung	0.960709	f17 20
liothek	0.960709	f17 20
htigung	0.960709	f17 20
roffene	0.960709	m7 20
gigkeit	0.960709	f17 20
ich	0.960694	m1 254
		n20 8
erk	0.960531	m1 1
		n20 56
ädel	0.960425	m4 22
rtel	0.960425	n23 22
ript	0.960425	n20 22
eher	0.960425	m4 22
akal	0.960425	m1 22
thek	0.960425	f17 22
ktur	0.960425	f17 22
ginal	0.960385	n20 21
ebung	0.960385	f17 21
sheit	0.960385	f17 21
ntage	0.960385	f16 21

erial	0.960385	n31 21
rling	0.960385	m1 21
eider	0.960385	m4 21
kurist	0.959749	m8 20
egramm	0.959749	n20 20
ektion	0.959749	f17 20
rblick	0.959749	m1 20
offene	0.959749	m7 20
iothek	0.959749	f17 20
ie	0.959696	m6 1
		m7 1
		f15 2
		f16 939
		n24 30
zip	0.959385	n31 23
igt	0.959385	f17 23
äre	0.959385	f16 23
kon	0.959385	m1 23
ukt	0.959385	n20 23
äß	0.958902	n20 26
esetzte	0.958687	f16 19
ulation	0.958687	f17 19
oration	0.958687	f17 19
gewicht	0.958687	n20 19
haftler	0.958687	m4 19
nachten	0.958687	n23 19
tiative	0.958687	f16 19
kennung	0.958687	f17 19
trum	0.958573	n28 21
tieg	0.958573	m1 21
zone	0.958573	f16 21
rial	0.958573	n31 21
flug	0.958573	m2 21
reis	0.958573	m1 21
dell	0.958573	n20 21
bchen	0.958447	n23 20
asche	0.958447	f16 20
ieher	0.958447	m4 20
welle	0.958447	f16 20
goner	0.958447	m4 20
osung	0.958447	f17 20
alist	0.958447	m8 20
ffene	0.958447	m7 20
rmung	0.958447	f17 20
urist	0.958447	m8 20
lette	0.958447	f16 20
tte	0.958181	m7 5
		f16 161
iative	0.957676	f16 19
setzte	0.957676	f16 19
aftler	0.957676	m4 19
öpfung	0.957676	f17 19
enieur	0.957676	m1 19
stätte	0.957676	f16 19
fläche	0.957676	f16 19
ohnung	0.957676	f17 19
ropole	0.957676	f16 19
kal	0.957561	m1 22
irr	0.957561	n20 22
eld	0.957561	n21 22
ipt	0.957561	n20 22
hek	0.957561	f17 22
min	0.957561	m1 22
ister	0.957351	m4 125
		n23 4
ktor	0.957231	m8 1
		m9 49
fene	0.956543	m7 20
zlei	0.956543	f17 20
denz	0.956543	f17 20
wamm	0.956543	m2 20
letzung	0.956441	f17 18
rfläche	0.956441	f16 18
ination	0.956441	f17 18
opole	0.956304	f16 19
nieur	0.956304	m1 19
ftler	0.956304	m4 19
etzte	0.956304	f16 19
igion	0.956304	f17 19
läche	0.956304	f16 19
tätte	0.956304	f16 19
ektor	0.955922	m8 1
		m9 46
rum	0.95557	n28 21
ieg	0.95557	m1 21
mie	0.95557	f16 21
iß	0.955509	m1 24
este	0.95548	m7 1
		f16 47
itzung	0.955372	f17 18
ission	0.955372	f17 18
nkunft	0.955372	f14 18
egerin	0.955372	f18 18
ahlung	0.955372	f17 18
orgnis	0.955372	f13 18
hester	0.955372	n23 18
kation	0.955372	f17 18
tzte	0.954297	f16 19
pole	0.954297	f16 19
ätte	0.954297	f16 19
pper	0.954297	m4 19
gion	0.954297	f17 19
lese	0.954297	f16 19
mune	0.954297	f16 19
lner	0.954297	m4 19
orie	0.954297	f16 19
dium	0.954297	n28 19
ieur	0.954297	m1 19
and	0.954272	m1 9
		m2 1041
		m3 2
		f14 25
		n22 9
u	0.954089	m1 2
		f17 1461
		n20 1
		n24 54
kurrenz	0.953944	f17 17
tretung	0.953944	f17 17
chauung	0.953944	f17 17

ndigung	0.953944	f17 17
elpunkt	0.953944	m1 17
hterung	0.953944	f17 17
ttung	0.953922	f17 18
nraum	0.953922	m2 18
rgnis	0.953922	f13 18
alter	0.953922	m4 18
retär	0.953922	m1 18
eifen	0.953922	m4 18
rtier	0.953922	n20 18
auung	0.953922	f17 18
nsion	0.953922	f17 18
onist	0.953922	m8 18
ip	0.953599	n31 23
gt	0.953599	f17 23
ald	0.953389	m3 20
üt	0.95295	n21 103
		n24 3
errede	0.95281	f16 17
retung	0.95281	f17 17
emonie	0.95281	f16 17
ichung	0.95281	f17 17
hauung	0.95281	f17 17
urrenz	0.95281	f17 17
räbnis	0.95281	n27 17
hkomme	0.95281	m7 17
lpunkt	0.95281	m1 17
tasche	0.95281	f16 17
tanz	0.951801	f17 18
etär	0.951801	m1 18
ifen	0.951801	m4 18
efon	0.951801	n20 18
trät	0.951801	n24 18
ände	0.951801	n23 18
mied	0.951801	m1 18
ölbe	0.951801	n23 18
ek	0.951507	f17 22
rrede	0.951273	f16 17
komme	0.951273	m7 17
uktur	0.951273	f17 17
riere	0.951273	f16 17
ndort	0.951273	m1 17
usion	0.951273	f17 17
nrohr	0.951273	n20 17
etung	0.951273	f17 17
stoff	0.951273	m1 17
nwand	0.951273	f14 17
erlei	0.951273	n24 17
dchen	0.951273	n23 17
äbnis	0.951273	n27 17
lzeit	0.951273	f17 17
sität	0.951273	f17 17
rrenz	0.951273	f17 17
eutsche	0.95114	f16 16
tnehmer	0.95114	m4 16
lligung	0.95114	f17 16
ftasche	0.95114	f16 16
mission	0.95114	f17 16
bachter	0.95114	m4 16
bkosung	0.95114	f17 16
duktion	0.95114	f17 16
wirkung	0.95114	f17 16
zierung	0.95114	f17 16
etation	0.95114	f17 16
ikation	0.95114	f17 16
lheimer	0.95114	m4 16
osphäre	0.95114	f16 16
tüm	0.950974	n20 19
tär	0.950974	m1 19
zte	0.950974	f16 19
üst	0.950974	n20 19
irkung	0.949934	f17 16
chelei	0.949934	f17 16
ension	0.949934	f17 16
allene	0.949934	m7 16
henend	0.949934	n25 16
bkreis	0.949934	m1 16
straße	0.949934	f16 16
itztum	0.949934	n22 16
sphäre	0.949934	f16 16
utsche	0.949934	f16 16
ublade	0.949934	f16 16
kosung	0.949934	f17 16
rain	0.949233	m6 1
		n24 41
po	0.949223	n24 21
bmal	0.949025	n22 17
dine	0.949025	f16 17
rohr	0.949025	n20 17
tant	0.949025	m8 17
trie	0.949025	f16 17
dort	0.949025	m1 17
kurs	0.949025	m1 17
olle	0.949025	f16 17
toff	0.949025	m1 17
stab	0.949025	m2 17
rgie	0.949025	f16 17
rlei	0.949025	n24 17
hen	0.948863	m4 61
		n23 1236
ang	0.948609	m2 863
		m6 42
nerin	0.948298	f18 16
traße	0.948298	f16 16
ehrer	0.948298	m4 16
blade	0.948298	f16 16
abend	0.948298	m1 16
nzeit	0.948298	f17 16
titut	0.948298	n20 16
elung	0.948298	f17 16
kreis	0.948298	m1 16
tztum	0.948298	n22 16
phäre	0.948298	f16 16
tinkt	0.948298	m1 16
enend	0.948298	n25 16
lange	0.948298	f16 16
herin	0.948298	f18 16
helei	0.948298	f17 16
llene	0.948298	m7 16
blo	0.948288	m6 18

fon	0.948288	n20 18
hör	0.948288	n20 18
urs	0.948288	m1 18
sserung	0.947971	f17 15
tierung	0.947971	f17 15
chwerde	0.947971	f16 15
klärung	0.947971	f17 15
ugtuung	0.947971	f17 15
uschung	0.947971	f17 15
sterung	0.947971	f17 15
ichelei	0.947971	f17 15
ersität	0.947971	f17 15
ömmling	0.947971	m1 15
henzeit	0.947971	f17 15
ndstein	0.947971	m1 15
nbacher	0.947971	m4 15
ift	0.947135	m1 3
		f17 126
		n20 2
ain	0.946683	m6 1
		n24 41
gtuung	0.946682	f17 15
hwerde	0.946682	f16 15
recher	0.946682	m4 15
enzeit	0.946682	f17 15
dstein	0.946682	m1 15
hbarin	0.946682	f18 15
nsport	0.946682	m1 15
rsität	0.946682	f17 15
erchen	0.946682	n23 15
serung	0.946682	f17 15
mmling	0.946682	m1 15
lust	0.946167	m1 80
		f14 3
vinz	0.945906	f17 16
ille	0.945906	f16 16
häre	0.945906	f16 16
nend	0.945906	n25 16
itte	0.945906	f16 16
lene	0.945906	m7 16
ange	0.945906	f16 16
inkt	0.945906	m1 16
itut	0.945906	n20 16
bend	0.945906	m1 16
ator	0.945906	m9 16
geld	0.945906	n21 16
eben	0.945906	n23 16
ztum	0.945906	n22 16
raße	0.945906	f16 16
wung	0.945906	m2 16
haus	0.945906	n22 16
eter	0.945521	m4 79
		n23 3
ohr	0.945301	n20 17
log	0.945301	m1 17
tab	0.945301	m2 17
chs	0.945301	m2 17
off	0.945301	m1 17
rasse	0.944935	f16 15
sport	0.944935	m1 15
wuchs	0.944935	m2 15
leife	0.944935	f16 15
lerei	0.944935	f17 15
tuung	0.944935	f17 15
chnam	0.944935	m1 15
nzung	0.944935	f17 15
mling	0.944935	m1 15
werde	0.944935	f16 15
gfrau	0.944935	f17 15
rchen	0.944935	n23 15
fchen	0.944935	n23 15
edung	0.944935	f17 15
barin	0.944935	f18 15
didat	0.944935	m8 15
rnative	0.944351	f16 14
chslung	0.944351	f17 14
enkunft	0.944351	f14 14
ordnung	0.944351	f17 14
ntation	0.944351	f17 14
dfinder	0.944351	m4 14
erne	0.943996	m7 1
		f16 37
üm	0.943946	n20 19
we	0.943946	f16 19
native	0.942967	f16 14
ponist	0.942967	m8 14
esende	0.942967	f16 14
törung	0.942967	f17 14
hslung	0.942967	f17 14
ffnung	0.942967	f17 14
fasser	0.942967	m4 14
finder	0.942967	m4 14
so	0.942636	m6 1
		n24 42
ient	0.942379	m8 15
chof	0.942379	m2 15
kweg	0.942379	m1 15
uchs	0.942379	m2 15
hnam	0.942379	m1 15
kier	0.942379	m6 15
dner	0.942379	m4 15
zahl	0.942379	f17 15
port	0.942379	m1 15
eite	0.942379	f16 15
ndal	0.942379	m1 15
idat	0.942379	m8 15
erde	0.942379	f16 15
form	0.942379	f17 15
arin	0.942379	f18 15
nade	0.942379	f16 15
eik	0.941943	m6 16
tut	0.941943	n20 16
orm	0.941943	f17 16
inz	0.941943	f17 16
ase	0.941943	f16 16
rne	0.941179	m7 1
		f16 37
ionär	0.94109	m1 14
iegen	0.94109	n23 14
adies	0.94109	n20 14

hteil	0.94109	m1 14
seite	0.94109	f16 14
platz	0.94109	m2 14
inken	0.94109	m4 14
assin	0.94109	m6 14
slung	0.94109	f17 14
ör	0.94086	n20 18
rs	0.94086	m1 18
rvative	0.940196	f16 13
waltung	0.940196	f17 13
brecher	0.940196	m4 13
strophe	0.940196	f16 13
störung	0.940196	f17 13
htvater	0.940196	m5 13
uchtung	0.940196	f17 13
terchen	0.940196	n23 13
rchtung	0.940196	f17 13
efinger	0.940196	m4 13
nalität	0.940196	f17 13
kschaft	0.940196	f17 13
tenheit	0.940196	f17 13
ole	0.939487	m7 2
		f16 55
ächter	0.93899	m4 1
		n23 32
ttform	0.938701	f17 13
vative	0.938701	f16 13
trophe	0.938701	f16 13
enraum	0.938701	m2 13
tvater	0.938701	m5 13
nabend	0.938701	m1 13
ittler	0.938701	m4 13
aktion	0.938701	f17 13
finger	0.938701	m4 13
ention	0.938701	f17 13
stunde	0.938701	f16 13
lösung	0.938701	f17 13
dstück	0.938701	n20 13
ntität	0.938701	f17 13
latz	0.938346	m2 14
läfe	0.938346	f16 14
lett	0.938346	n20 14
onär	0.938346	m1 14
egen	0.938346	n23 14
eppe	0.938346	f16 14
auch	0.938346	m2 14
dies	0.938346	n20 14
nam	0.938145	m1 15
amt	0.938145	n22 15
dal	0.938145	m1 15
und	0.937919	m1 1211
		m2 68
		m3 6
ab	0.937427	m2 17
og	0.937427	m1 17
am	0.937427	m1 17
ntion	0.936676	f17 13
rette	0.936676	f16 13
rzeug	0.936676	n20 13
pchen	0.936676	n23 13
teige	0.936676	f16 13
leppe	0.936676	f16 13
rfluß	0.936676	m2 13
rophe	0.936676	f16 13
tunde	0.936676	f16 13
telle	0.936676	f16 13
ttler	0.936676	m4 13
enade	0.936676	f16 13
tität	0.936676	f17 13
ösung	0.936676	f17 13
derei	0.936676	f17 13
ehren	0.936676	n23 13
kampf	0.936676	m2 13
tform	0.936676	f17 13
ze	0.935461	m7 38
		f16 650
		n23 1
egel	0.935412	m4 66
		f16 3
edition	0.935351	f17 12
lendung	0.935351	f17 12
tattung	0.935351	f17 12
pektive	0.935351	f16 12
dwerker	0.935351	m4 12
städter	0.935351	m4 12
htigall	0.935351	f17 12
zeichen	0.935351	n23 12
tration	0.935351	f17 12
ielerin	0.935351	f18 12
iterung	0.935351	f17 12
ot	0.93516	m6 1
		n20 175
		n24 8
lege	0.933912	m7 48
		f16 2
uze	0.933798	f16 14
ies	0.933798	n20 14
eau	0.933798	n24 14
äfe	0.933798	f16 14
när	0.933798	m1 14
igte	0.933729	m7 31
		f16 1
stelle	0.933727	f16 12
werker	0.933727	m4 12
ätzung	0.933727	f17 12
uldner	0.933727	m4 12
rüßung	0.933727	f17 12
tigall	0.933727	f17 12
rzehnt	0.933727	n20 12
tädter	0.933727	m4 12
ündung	0.933727	f17 12
dtchen	0.933727	n23 12
ferenz	0.933727	f17 12
ikaner	0.933727	m4 12
ikatur	0.933727	f17 12
ektive	0.933727	f16 12
arette	0.933727	f16 12
trolle	0.933727	f16 12
eldung	0.933727	f17 12
attung	0.933727	f17 12

affung	0.933727	f17 12
kplatz	0.933727	m2 12
htisch	0.933727	m1 12
ücke	0.933713	f16 13
rang	0.933713	m2 13
kopf	0.933713	m2 13
ampf	0.933713	m2 13
ränk	0.933713	n20 13
rker	0.933713	m4 13
inar	0.933713	n20 13
wanz	0.933713	m2 13
zern	0.933713	m1 13
ldner	0.931525	m4 12
üßung	0.931525	f17 12
kaner	0.931525	m4 12
chirr	0.931525	n20 12
chler	0.931525	m4 12
katur	0.931525	f17 12
ffung	0.931525	f17 12
igall	0.931525	f17 12
ktive	0.931525	f16 12
nbild	0.931525	n21 12
erker	0.931525	m4 12
erenz	0.931525	f17 12
ähler	0.931525	m4 12
rität	0.931525	f17 12
nfall	0.931525	m2 12
ublik	0.931525	f17 12
lling	0.931525	m1 12
zerin	0.931525	f18 12
schuß	0.931525	m2 12
ption	0.931525	f17 12
rolle	0.931525	f16 12
ädter	0.931525	m4 12
aille	0.931525	f16 12
zehnt	0.931525	n20 12
rtler	0.931525	m4 12
eich	0.931501	m1 138
		n20 8
mme	0.931266	m7 26
		f16 405
nde	0.930244	m7 43
		f16 889
		n23 18
bredung	0.929659	f17 11
enleben	0.929659	n23 11
söhnung	0.929659	f17 11
euerung	0.929659	f17 11
rsacher	0.929659	m4 11
hfolger	0.929659	m4 11
atmeter	0.929659	m4 11
uigkeit	0.929659	f17 11
igerung	0.929659	f17 11
trat	0.929426	m1 29
		n20 1
mt	0.929194	n22 15
hiv	0.928804	n20 13
änk	0.928804	n20 13
gon	0.928804	m6 13
nur	0.928804	f14 13
maß	0.928804	n20 13
oge	0.928804	m7 13
ern	0.928804	m1 13
irk	0.928804	m1 13
blik	0.928304	f17 12
ücht	0.928304	n20 12
kzug	0.928304	m2 12
phie	0.928304	f16 12
ssur	0.928304	f17 12
übde	0.928304	n23 12
rone	0.928304	f16 12
dter	0.928304	m4 12
stin	0.928304	f18 12
rale	0.928304	f16 12
etit	0.928304	m1 12
iser	0.928304	m4 12
onna	0.928304	f15a 12
hirr	0.928304	n20 12
nnig	0.928304	m1 12
ehnt	0.928304	n20 12
gall	0.928304	f17 12
epunkt	0.92788	m1 11
tmeter	0.92788	m4 11
nleben	0.92788	n23 11
ttchen	0.92788	n23 11
uerung	0.92788	f17 11
hdruck	0.92788	m1 11
hwuchs	0.92788	m2 11
elchen	0.92788	n23 11
redung	0.92788	f17 11
sacher	0.92788	m4 11
achung	0.92788	f17 11
pathie	0.92788	f16 11
öhnung	0.92788	f17 11
baller	0.92788	m4 11
ßstadt	0.92788	f14 11
ustrie	0.92788	f16 11
athie	0.925469	f16 11
izist	0.925469	m8 11
ratur	0.925469	f17 11
miede	0.925469	f16 11
strie	0.925469	f16 11
aller	0.925469	m4 11
rgeld	0.925469	n21 11
öpfer	0.925469	m4 11
ndler	0.925469	m4 11
zisse	0.925469	f16 11
leben	0.925469	n23 11
hafen	0.925469	m5 11
runde	0.925469	f16 11
hjahr	0.925469	n20 11
reuer	0.925469	m4 11
traum	0.925469	m2 11
trakt	0.925469	m1 11
eiung	0.925469	f17 11
elier	0.925469	m1 11
rzahl	0.925469	f17 11
ren	0.925193	m4 4
		n23 73
ack	0.924446	m2 1

		m3 58	ndlage	0.920866	f16 10
		n24 2	source	0.920866	f16 10
es	0.924185	n20 14	weiche	0.920866	f16 10
go	0.924185	m6 14	orität	0.920866	f17 10
na	0.924185	f15a 14	utzung	0.920866	f17 10
macht	0.924174	f14 1	uptung	0.920866	f17 10
		f17 26	ellenz	0.920866	f17 10
nder	0.923571	m4 96	ichter	0.920866	m4 10
		n23 6	lehrer	0.920866	m4 10
me	0.923502	m7 51	mtheit	0.920866	f17 10
		f16 798	nister	0.920866	m4 10
		n23 8	menade	0.920866	f16 10
ett	0.923215	n20 57	iterin	0.920866	f18 10
		n24 1	aurant	0.920866	n24 10
		n25 2	berger	0.920866	m4 10
nna	0.922968	f15a 12	bewerb	0.920866	m1 10
lon	0.922968	m6 12	uldung	0.920866	f17 10
rik	0.922968	f17 12	hützer	0.920866	m4 10
nig	0.922968	m1 12	tzbube	0.920866	m7 10
ail	0.922968	n24 12	rkunft	0.920866	f14 10
tue	0.922968	f16 12	schein	0.920866	m1 10
bde	0.922968	n23 12	idiger	0.920866	m4 10
hnt	0.922968	n20 12	enfall	0.920866	m2 10
ber	0.922848	m4 281	ghafen	0.920866	m5 10
		n23 20	asse	0.919734	m7 2
ütigung	0.922833	f17 10			f16 39
ichwort	0.922833	n22 10	merei	0.918199	f17 10
auptung	0.922833	f17 10	lwerk	0.918199	n20 10
erkunft	0.922833	f14 10	recht	0.918199	n20 10
richter	0.922833	m4 10	kodil	0.918199	n20 10
ezimmer	0.922833	n23 10	chein	0.918199	m1 10
iligung	0.922833	f17 10	kasse	0.918199	f16 10
eidiger	0.922833	m4 10	hrede	0.918199	f16 10
beitung	0.922833	f17 10	enkel	0.918199	m4 10
eiterin	0.922833	f18 10	ource	0.918199	f16 10
henfall	0.922833	m2 10	licht	0.918199	n21 10
chützer	0.922833	m4 10	immel	0.918199	m4 10
taurant	0.922833	n24 10	mplar	0.918199	n20 10
tstätte	0.922833	f16 10	tgang	0.918199	m2 10
huldung	0.922833	f17 10	rdung	0.918199	f17 10
techung	0.922833	f17 10	pferd	0.918199	n20 10
tbewerb	0.922833	m1 10	ptung	0.918199	f17 10
sheimer	0.922833	m4 10	urant	0.918199	n24 10
hode	0.921942	f16 11	erger	0.918199	m4 10
luck	0.921942	m1 11	eiche	0.918199	f16 10
afen	0.921942	m5 11	amide	0.918199	f16 10
kind	0.921942	n21 11	nomie	0.918199	f16 10
rakt	0.921942	m1 11	llenz	0.918199	f17 10
iung	0.921942	f17 11	idenz	0.918199	f17 10
zist	0.921942	m8 11	öhung	0.918199	f17 10
erve	0.921942	f16 11	ewerb	0.918199	m1 10
iede	0.921942	f16 11	derin	0.918199	f18 10
bart	0.921942	m2 11	trast	0.918199	m1 10
hnik	0.921942	f17 11	ägung	0.918199	f17 10
alle	0.921942	f16 11	dlage	0.918199	f16 10
rurg	0.921942	m8 11	zbube	0.918199	m7 10
thie	0.921942	f16 11	demie	0.918199	f16 10
emie	0.921942	f16 11	emble	0.918199	n24 10
lag	0.92145	m1 19	ützer	0.918199	m4 10
		m2 263	nomen	0.918199	n20 10

wort	0.91787	f17 246
		n20 5
		n22 14
bar	0.916662	m7 65
		m8 4
dit	0.9161	m1 11
rve	0.9161	f16 11
üre	0.9161	f16 11
gge	0.9161	f16 11
end	0.915586	m1 16
		f17 485
		n20 8
		n25 16
ve	0.915276	m7 6
		f16 94
rent	0.914299	m8 10
omen	0.914299	n20 10
bube	0.914299	m7 10
lyse	0.914299	f16 10
mble	0.914299	n24 10
plar	0.914299	n20 10
ativ	0.914299	n20 10
urce	0.914299	f16 10
odil	0.914299	n20 10
lenz	0.914299	f17 10
hein	0.914299	m1 10
mide	0.914299	f16 10
nche	0.914299	f16 10
ferd	0.914299	n20 10
mant	0.914299	m8 10
werb	0.914299	m1 10
rast	0.914299	m1 10
buch	0.914299	n22 10
omie	0.914299	f16 10
mium	0.914299	n28 10

dat	0.912943	m8 62
		n20 4
lucht	0.912129	f14 4
		f17 58
uße	0.911442	m7 24
		f16 1
nke	0.910744	m7 435
		f16 38
osse	0.909625	m7 46
		f16 3
to	0.909465	m6 1
		n24 26
mel	0.908807	m4 245
		f16 21
äck	0.907837	n20 10
rch	0.907837	m2 10
tom	0.907837	n20 10
dil	0.907837	n20 10
yse	0.907837	f16 10
rce	0.907837	f16 10
äut	0.907837	n20 10
erd	0.907837	n20 10
ord	0.907837	m1 10
bund	0.906095	m1 66
		m2 5
io	0.903747	n24 11
dy	0.903747	m6 11
ös	0.903747	m1 11
if	0.903747	m1 11
ße	0.903647	m7 24
		f16 270
oß	0.902593	m1 1
		m2 28
		n22 317