Guessing Morphological Classes of Unknown German Nouns

Preslav Nakov[1], Yury Bonev[2], Galia Angelova[3], Evelyn Cius[4], Walther von Hahn[4]

(1) University of California at Berkeley, EECS, Berkeley CA 94720, USA
(2) Sofia University "St. Kl. Ohridski" and Team Vision Bulgaria
(3) Central Laboratory for Parallel Processing, Bulgarian Academy of Sciences, 25 Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
(4) University of Hamburg, CS Dept., Vogt-Koelln Str.-30, 22527 Hamburg, Germany

nakov@eecs.berkeley.edu, y.bonev@team-vision.bg, galia@lml.bas.bg, evelyn.gius@imail.de, vhahn@informatik.uni-hamburg.de

Abstract

A system for recognition and morphological classification of unknown German words is described. Given raw texts it outputs a list of the unknown nouns together with hypotheses about their possible stems and morphological class(es). The system exploits both global and local information as well as morphological properties and external linguistic knowledge sources. It learns and applies ending-guessing rules similar to the ones originally proposed for POS guessing. The paper presents the system design and implementation and discusses its performance by extensive evaluation. Similar ideas for ending-guessing rules have been applied to Bulgarian as well but the performance is worse due to the difficulties of noun recognition as well as to the highly inflexional morphology with numerous ambiguous endings.

1 Introduction

The recognition and relevant processing of unknown words is a primary problem for each Natural Language Processing (NLP) system. No matter how big lexicon it has, it always meets unknown wordforms in the real texts as new words are constantly added to the language. Linguistic phenomena such as *derivation*, *compounding* and *inflection* (to some extent), constantly generate new wordforms, new *proper names* appear and *foreign words* are adopted on a daily basis[1]. The majority of the current systems either use spell-checkers, lists of exceptions and gazetteers of proper names to support the recognition of strings that look like words but do not appear in the system's lexicons or rely on data-driven approaches in order to model the encountered phenomena and decide on the type and category of the new wordforms met in the text.

While the majority of the available systems for automatic processing of unknown words aim at the recognition of the most probable *Part Of Speech* (*POS*) tag, our system called **MorphoClass**[2] tries to guess the *morphological class* of unknown words likely to be *nouns* in a certain text. It identifies the unknown German wordforms, "gathers" them into groups as candidates belonging to a single paradigm and attempts to propose both a suitable *stem* and a *morphological class*. To the best of our knowledge, **MorphoClass** is the only system that attempts to address these morphological issues.

We define the *stem* as the common part shared by all inflected wordforms (up to valid alternations). Together with the morphological class it determines unambiguously all the wordforms that could be obtained through inflection of the same base paradigm. **MorphoClass** is as a kind of a tool for lexical acquisition: it identifies new wordforms in the raw text, derives some properties and performs morphological classification. It can be used as a tool for automatic dictionary extension with new words.

MorphoClass solves the "guessing" problem as a sequence of subtasks including:

- identification of unknown words (at present, limited to nouns only);
- recognition and grouping of the inflected forms of the same word (they share the same stem);
- compound splitting;
- morphological stem analysis;
- stem hypothesis generation for each group of inflected forms;
- ranking the list of hypotheses about the possible morphological class for each group of words.

[1] Misspelled words as well as orthographic variants resulting from e.g. the recent reform of the German orthography are also "new words" for systems performing automatic text analysis.

[2] **MorphoClass** was developed within the EC funded project "BIS-21 Centre of Excellence" ICA1-2000-70016 and was additionally supported by the bilateral cooperation programme between the Hamburg University and the Sofia University "St. Kl. Ohridski".

This is a several-stage process, which relies on:

- **morphology** – compound splitting, inflection, affixes;
- **global context** – wordforms collected from the whole input, word frequency statistics, ending-guessing rules etc.**;**
- **local context** – surrounding words: articles, prepositions, pronouns;
- **external sources** – especially designed lexicons, German grammar information etc.

MorphoClass is not a POS guesser in the traditional sense. The purpose of the latter is to propose a hypothesis about the possible POS of an unknown word by observing its graphemic form in the particular local context and possibly in a lexicon. But **MorphoClass** is not restricted to the local context: it collects and considers all the word occurrences throughout the whole input, trying to identify groups of inflectional forms of the same word and to derive a hypothesis for the corresponding morphological class. **MorphoClass** as a kind of **morphological class guesser** might work after a POS tagger completes its tasks and tags the unknown nouns (but it can be run *before* the tagger as well and thus support its decisions). **MorphoClass** can also be used as a lemmatiser, since it outputs both the stem and the morphological class for each known word. At the same time **MorphoClass** is not a stemmer in the classic Information Retrieval (IR) sense since it does not conflate the derivational forms: e.g. *generate* and *generator* would be grouped together by most of the IR stemmers but not by **MorphoClass**.

The paper is organised as follows. Section 2 sketches related work. Section 3 presents the system resources and architecture. Section 4 discusses the **MorphoClass** ending-guessing rules in more detail. Section 5 presents the system at work. Section 6 contains the evaluation, and section 7 – further improvements by linear context consideration. Section 8 contains the conclusion and future work.

2 Related Work

We studied numerous relevant papers looking for insights that might be useful especially in the evaluation of **MorphoClass**. But, as we mentioned above, we are not aware of any other system that tries to guess the morphological class by observing the endings only (after collecting the relevant wordforms spread in the text) and without considering any word formation rules. Still, our approach is more or less related to and influenced by several classical NLP tasks, the nearest ones being morphological analysis and POS tagging. Below we list some related work putting the emphasis on German compounds splitting and guessing rules for POS recognition.

German morphology. Finkler and Neumann use *n*-ary tries in the *MORPHIX* system (Finkler 88). (Adda-Decker & Adda 00) propose rules for morpheme boundary identification: after the occurrence of some sequences like: *-ungs*, *-hafts*, *-lings*, *-tions*, *-heits*. (Neumann 99) considers the problem of compound analysis by means of longest matching substrings found in a lexicon. The problem of German compound splitting is considered in depth by (Goldsmith 98; Lezius 00; Ulmann 95; Hietsch 84). The latter concentrates on the function of the last part of the compound. Recent projects such as DeKo focus on the collection of word formation rules, lexicons and software tools for the analysis of complex German words (DeKo 01).

POS guessing. (Kupiec 92) uses pre-specified suffixes and performs statistical learning for POS guessing. The XEROX tagger comes with a list of built-in ending-guessing rules (Cutting et al. 92). In addition to the ending (Weischedel et al. 93) exploit the capitalisation. (Thede & Harper 97) consider contextual information, word endings, entropy and open-class smoothing. A similar approach is presented in (Schmid 95). A very influential is the work of Brill (Brill 97) who builds more linguistically motivated rules exploiting both a tagged corpus and a lexicon. He does not look at the affixes only but also checks their POS class in a lexicon. Mikheev proposes a similar approach but learns the ending-guessing rules from a *raw* as opposed to *tagged* text (Mikheev 97). (Daciuk 99) speeds up the process by means of finite state transducers.

General morphology. Schone and Jurafsky use latent semantic analysis for a knowledge-free morphology induction (Schone & Jurafsky 00). (Goldsmith 01) performs a minimum description length analysis of the morphology of several European languages using corpora. (DeJean 98) and (Hafer & Weiss 74) follow a successor variety approach: the word is cut if the number of distinct letters after a pre-specified sequence surpasses some threshold. (Gaussier 99) induces derivational morphology from a lexicon by means of splitting based on *p*-similarity. (Jacquemin 97) focuses on the morphological processes. (Van den Bosch & Daelemans 99) propose a memory-based approach mapping directly from letters in context to categories encoding morphological boundaries, syntactic class labels and spelling changes. (Yarowsky & Wicentowski 00) present a corpus-based approach for morphological analysis of both regular and irregular forms based on four models including: relative corpus frequency, context similarity, weighted string similarity and incremental retraining of inflectional transduction probabilities. Another interesting work exploiting capitalisation as well as fixed and variable suffixes is presented in (Cucerzan & Yarowsky 00).

3 MorphoClass Resources and Architecture

Figure 1 shows the **MorphoClass** linguistic resources used and the main modules architecture. The *Stem Lexicon* (*SL*) is compiled from both the NEGRA corpus (NEGRA 01) and the full-form Morphy lexicon (Lezius 00) and currently contains 13,147 German nouns (words like *der/die/das Halfter* are represented as three lexicon items with a separate morphological class each). SL facilitates the compounds recognition since the compound splitting module relies on the noun stems it contains. The *Expanded Stem Lexicon* (*ESL*) includes all wordforms derived from the SL entries and has been used substantially during the ending-guessing rules learning stage. The *Word Lexicon* (*WL*) contains important closed-class words such as articles, pronouns, prepositions etc., which are often met in the text as part of the local context surrounding the unknown words.

Class	Singular				Plural			
	Nom	Gen	Dat	Akk	Nom	Gen	Dat	Akk
m1	0	[e]s(1)	[e]	0	e	e	en	e
…	…	…	…	…	…	…	…	…
m3a	0	[e]s(1)	[e]	0	er	er	ern	er
…	…	…	…	…	…	…	…	…
m9	0	[e]s(1)	[e]	0	en	en	en	en
…	…	…	…	…	…	…	…	…
n20	0	[e]s(1)	[e]	0	e	e	en	e
n21	0	[e]s(1)	[e]	0	er	er	ern	er
…	…	…	…	…	…	…	…	…
n25	0	[e]s(1)	[e]	0	en	en	en	en

Table 1: Morphological classes of German nouns

The **MorphoClass** morphological classes have been designed for the DB-MAT system (DB/R/-MAT 92-98); we reduced the original 41 classes to 39, which are not sensitive to stress alternation. Each class contains 8 noun endings (see Table 1), which are to be augmented to the stem with possible alternations and other changes expressed by rules. For instance, the ending *–s* for *Genetive Singular* is augmented to stems from class **m1** after adding a preceding *–e-* and taking into account *rule 1*, which includes statements like: "when the basic nominative form ends by s/sch/x/chs/z/tz/… the vowel *–e-* is obligatory". A complete list of the morphological classes and rules is presented in (Nakov et al. 02).

Figure 1 sketches the sequence of tasks for the identification and morphological classification of unknown words but in what follows we will focus on the processing of unknown *nouns* only. The successful recognition of unknown nouns substantially depends on the fact that the German nouns are capitalised (so each capitalised word from the text is considered as either a noun, initial sentence word or named entity).

MorphoClass outputs one of the following three kinds of indications for each group of wordforms sharing a common stem:

- COMPOUND – successfully split using the available lexicon, so the morphological class of the last word in the compound is assigned;
- ENDING RULE – an ending-guessing rule has been applied and the predicted class is assigned;
- NO INFO – no decision was taken beyond the incompatible classes elimination.

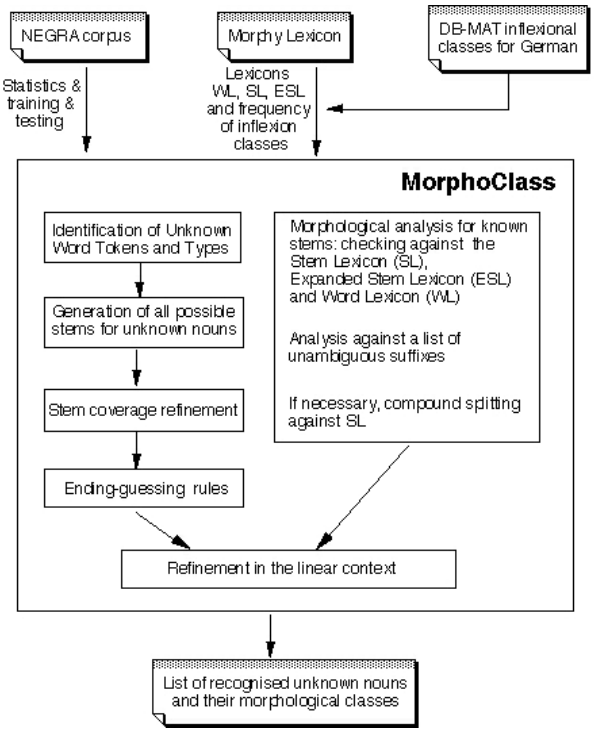

Figure 1: System resources and main modules.

4 Ending-Guessing Rules

For the morphological class prediction we adopted an ending-guessing rules mechanism which has been originally proposed for POS guessing (Mikheev 97). We built 482 rules when running the rule induction against the SL and 1,789 rules when the SL entries were weighted according to their frequencies in a raw text (see Table 2). We considered *all* endings up to 7 characters long and met at least 10 times in a raw training text, provided that there were at least 3 characters remaining to the left, including at least one vowel. For each noun we extracted all possible suffixes (e.g. from *Vater* we obtain *-r*, *-er* and *-ter*) and for each ending – a list of the morphological classes it appeared with and their corresponding frequencies. It is intuitively clear that a good ending-guessing rule would be: *unambiguous* (predicts a particular class without or with only few exceptions); *frequent* (must be based on a large number of

occurrences); *long* (the longer the ending, the less the probability that it will appear by chance, and thus the better its prediction).

Although the maximum likelihood estimation is a good predictor of the rule quality, it takes into account neither the rule length nor the rule frequency. So, following (Mikheev 97) we adopted the score:

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

$$p = (x+0.5)/(n+1),$$

where:

- l is the ending length;
- x is the number of successful rule guesses;
- n is the total number of training instances compatible with the rule;
- p is a modified version of the *maximum likelihood estimation* $\hat{p}$ which ensures that neither p nor $(1-p)$ could be zero;
- $\sqrt{\frac{p(1-p)}{n}}$ is an estimation of the dispersion;
- $t_{(1-\alpha)/2}^{(n-1)}$ is a coefficient of the t-distribution (with $(n-1)$ degrees of freedom and confidence level α).

Ending	Confidence	Predicted class(es)	Class frequency(s)
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

Table 2: The most confident ending-guessing rules (learned on the lexicon and weighted on a raw text).

We keep the rules whose score is above some threshold. Currently, we use 0.90 – a high level that guarantees we can rely on the rules' predictions. This makes the system conservative in a sense that no specific morphological class (out of the feasible ones for the target group of wordforms) is proposed unless **MorphoClass** is confident enough in the choice. If more texts are presented to the system it would possibly observe more wordforms of the target unknown noun and will rule out some of the possible classes. If we want a choice at any price, we can move the threshold down or allow ambiguous ending-guessing rules that predict more than one morphological class.

More on the ending-guessing rules, and a full list of the ones used, can be found in (Nakov et al. 02).

5 Examples

MorphoClass goes through the input text, collects unknown noun wordforms, groups them and attempts to generate a stem for each group, as shown in Table 3. For each stem in column one it checks whether there exists a morphological class that could generate all the wordforms listed in column three. If at least one is found, **MorphoClass** accepts the current coverage as currently feasible. Otherwise the system tries to refine it in order to make it acceptable. It is possible that the same stem is generated by a set of words that could not be covered together as members of the same paradigm. At the first step we are not interested whether the stem is really the correct one but just in whether it is *compatible* with all the wordforms it covers taken together, i.e. whether there exists a morphological class that could generate them all. For instance, if there is only one unknown wordform of a certain paradigm, e.g. *Tages*, all possible stems will be generated: *Tages*, *Tage* and *Tag*. All the three stems are valid since they have been obtained by reversing legal declination rules only. Stem refinement is possible after collecting more wordforms occurrences from the same paradigm.

Stem	#	Wordforms 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 }
Fall	5	{ Fall, Falle, Falles, Fälle, Fällen }
Geschäft	5	{ Geschäft, Geschäfte, Geschäften, Geschäftes, Geschäfts }
Schrei	3	{ Schrei, Schreien, Schreier}

Table 3: Largest "coverage" stems, ordered by the number of word types "covered".

How to refine the Table 3 rows? An obvious (but not very wise) solution is just to reject any stem that seems to cover "contradicting" wordforms. But we are not willing to do so since we might lose a useful stem. For example, we do not have to reject the stem *Spiel* just because it is incompatible with the set of words shown in Table 3. We have to decide that *Spiel*, *Spiele*, *Spielen* and *Spiels* are correct members of the *Spiel*-paradigm, while *Spieler* are *Spielern* are

not and probably belong to a different one. The first group of wordforms – *Spiel*, *Spiele*, *Spielen* and *Spiels* might be generated from *Spiel* by four classes, two masculine and two neutrum (*m1*, *m9*, *n20* and *n25*), while the second one – *Spieler*, *Spielern* – may be generated from *Spiel* by two classes, one masculine and one neutrum (*m3a* and *n21*, see Table 1). Thus, both groups are acceptable taken separately. The first group is bigger and therefore more likely to be correct. So we decide that the first four wordforms belong to the paradigm of *Spiel.* Applying ending-guessing rules, we will have to choose between four possible morphological classes (*m1*, *m9*, *n20* and *n25*). For *Spieler* and *Spielern* **MorphoClass** will continue to explore other possible stems. If the two groups of wordforms had the same number of members, we would have taken the most likely morphological class, which appeared more frequently according to the statistics collected from the Morphy's lexicon and NEGRA. In the worst case **MorphoClass** will propose two candidates.

What is important is that we *choose* between the two groups. By doing so we presuppose that the stem *Spiel* has *exactly one* morphological class. In fact it is relatively rear for a noun to have more than one class: SL contains just 73 such stems out of 13,147. It is even more unlikely that a new unknown word will have more than one morphological class, and additionally that such a new word is used with two or more of these classes in the same text. So, we always look for one paradigm only, preferring the biggest set that a morphological class could cover.

Table 4 is another illustration of the refinement algorithm. It lists the top unknown stems found in the NEGRA corpus ordered by the number of wordforms covered (and then alphabetically). Table 4 provides good examples for illustrating the interaction between the compound-splitting and ending-guessing modules. Let us consider the 3rd raw – *Bildungsurlaub* (educational holiday, study leave) which is a compound constructed by *Bildung* (study) and *Urlaub* (vacation, leave). The last noun *Urlaub* determines both the gender (masculine) and the morphological class of the whole compound. However, a plural form of this paradigm *Bildungsurlaube* (see the last raw of Table 4) covers the base form *Laube* (summer house), which is a feminine noun in German. The 3rd raw as well as the last one include the form *Bildungsurlauber* (person who is in study leave) which does not belong to these paradigms.

How does **MorphoClass** deal with all these strings of letter? As a first step, it identifies the allowed compounds using the nouns available in the system's lexicons plus some knowledge about the correct construction of German compounds. In this case, it will face the stem *Bildungsurlaub* first, as is suggested by the 4 wordforms, and will find that *Bildung* as well as *Urlaub* are included in the lexicon, i.e. *Bildung-s-urlaub* is a valid compound which however does not generate *Bildungsurlauber* as a member of the same paradigm. At this moment *Bildung-s-urlaub* will be remembered as a compound and *Bildungsurlauber* will be kept as a single form that may belong to another paradigm (this case was considered above in Table 3).

Unknown Stem	#	Words that Generated the Stem
Ortsbeirat	5	{ Ortsbeirat, Ortsbeirates, Ortsbeirats, Ortsbeiräte, Ortsbeiräten }
Bildungsurlaub	**4**	**{ Bildungsurlaub, Bildungsurlaube, Bildungsurlauben, Bildungsurlauber }**
Gemeindehaushalt	4	{ Gemeindehaushalt, Gemeinde-haushalte, 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 }
Bildungsurlaube	**3**	**{ Bildungsurlaube, Bildungsurlauben, Bildungsurlauber }**

Table 4: Unknown stems, ordered by the number of the covered word tokens.

Seeing *Bildungsurlaube* in the last raw of Table 4 **MorphoClass** will try to decompose it as *Bildungsur-laube* since *Laube* is in the lexicon too (but will not find *Bildungsur-* as a possible initial part for building a compound, moreover *Laube* has no form *Lauber* and thus the from *Bildungsurlauber* cannot be generated). If the decomposition *Bildungsur-laube* had been successful **MorphoClass** would had kept this stem with suggestion of another compound (again excluding *Bildungsurlauber* from the paradigm). But since the decomposition *Bildungsur-laube* fails, **MorphoClass** will consider the initial two forms at the last raw of Table 4 as belonging to the paradigm in the 3rd raw, as *Bildungsurlaub* is once treated as a valid component. If *Laube* is not in the lexicon, **MorphoClass** will not consider the possibility of its existence.

Now imagine that neither *Bildung*, *Urlaub* nor *Laube* were present in the lexicon. **MorphoClass** will try to apply ending-guessing rules to *Bildungsurlaub* at the 3rd raw, will determine some morphological class and will exclude *Bildungsurlauber* from the paradigm as an impossible member (it will be considered separately later). Reaching the last raw, **MorphoClass** will not consi-

der *Bildungsurlaube* as a possible stem, as it is already covered by a stem (namely *Bildungsurlaub*). In this way **MorphoClass** will not guess that *Bildung*, *Urlaub*, and *Laube* are possible base forms of German nouns.

6 Evaluation

The **MorphoClass** system has been manually evaluated over the following texts:

- **Kafka:** *Erzählungen* by Kafka, 3,510 word types, 13,793 word tokens;
- **Goethe 1:** *Die Wahlverwandtschaften* by Goethe, 10,833 word types, 79,485 word tokens;
- **Goethe 2:** *Wilhelm Meisters Lehrjahre* by Goethe, 17,252 word types, 194,266 word tokens (Goethe 2).

These electronic versions contained no misspelled words but if there were, **MorphoClass** would consider them as unknown and would try to group them and propose possible stem(s). Table 5 contains detailed statistics of the kinds of wordforms found by manual investigation of the different texts, after **MorphoClass** had suggested stems for the unknown nouns. Note, that these categories are overlapping, so that a particular word token can belong to more than one of them.

	not a noun	proper name	wrong stem	multiple class	singular only	plural only	**total**
Kafka	8 1.7%	13 2.77%	72 15.11%	26 5.32%	80 16.81%	3 0.64%	473 100%
Goethe 1	29 1.69%	22 1.28%	320 18.73%	194 11.49%	232 13.77%	11 0.64%	1,706 100%
Goethe 2	42 1.48%	106 3.74%	481 16.95%	201 7.08%	343 12.09%	20 0.70%	2,838 100%

Table 5: Kinds of wordforms – detailed statistics.

As we said above, **MorphoClass** considers some words as candidates for nouns (normally proper nouns and foreign words are included) and tries to decide what the corresponding inflectional class is. Sometimes the assignment is impossible (mostly when just one wordform is found) and the system indicates that there is no enough information of how to propose an inflectional class since neither the compound splitting nor the ending-guessing rules (with confidence above the threshold of 0.9) were applicable. This is a positive feature of **MorphoClass** since it avoids misleading decisions in the case of absent information. Table 6 summarises the **MorphoClass** reactions for the three testing data sets. We should emphasize that **MorphoClass** always produces a list of the feasible candidate classes but in the case of "no info" it does not prefer any of them.

The ending-guessing rules were applied only if the compound-splitting ones had 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. A substantial amount of the remaining stems are covered by the ending-guessing rules. Table 6 shows that the ending-guessing rules are applied for more than 40% of the stems (45% on the average). Their precision in isolation was much lower (see the details below). It should be noted however, that **MorphoClass** has no dictionary of named entities and that its ending-guessing rules were trained on 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-guessing rules applied at present are based on representative statistics about the possible endings of the German nouns. All results should be considered as relative, according to the available resources. No doubt, a list of named entities and a better initial lexicon would influence considerably the results presented.

	Nouns	Compo-ounds	Ending-guessing	"No info"-stems
Kafka	473	185	190 (40%)	98 (21%)
Goethe 1	1,706	551	837 (49%)	318 (19%)
Goethe 2	2,838	896	1,274 (45%)	668 (23%)

Table 6: Noun wordforms in the different texts.

The stems are split into the following categories:

- **SET:** a *set* of classes is assigned instead of a single one.
- **PART:** a *correct* class is discovered but *not all* the correct ones;
- **WRONG:** a single class is assigned but it is *wrong*;
- **YES:** a single class is assigned and it is the only correct one;
- **SKIP:** the stem has been excluded from the current manual evaluation (proper names, non-German nouns, non-nouns or wrong stem).

We define *precision* and *coverage* as follows:

$$\text{precision1} = \text{YES} / (\text{YES} + \text{WRONG} + \text{PART})$$
$$\text{precision2} = (\text{YES} + (\textit{scaled\_PART})) / (\text{YES} + \text{WRONG} + \text{PART})$$
$$\text{precision3} = (\text{YES} + \text{PART}) / (\text{YES} + \text{WRONG} + \text{PART})$$
$$\text{coverage} = (\text{YES} + \text{WRONG} + \text{PART}) / (\text{YES} + \text{WRONG} + \text{PART} + \text{SET})$$

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 g to the total classes count: if a stem belongs to k ($k > 2$) classes and **MorphoClass** found g of them (it finds exactly one in case of ending-guessing rule but in case of compound splitting or no

rule applicable, it can find more) *precision1* considers this as a failure (will add 0), *precision2* counts it as a partial success (will add *scaled_PART* = *g*/*k*) and *precision3* accepts it as a success (will add 1).

Table 7 shows the results of the manual evaluation using the measures defined above. As we can see, the bigger the text the higher the precision but less the coverage. In addition, we considered two baselines. *Baseline 1* always predicts class *f16* since it is the most frequent one. *Baseline 2* proposes the most frequent class but limited to the ones compatible with both the proposed stem and the wordform group. Tables 8.1 and 8.2 show the performance of **MorphoClass** is well above both baselines.

	Kafka	Goethe 1	Goethe 2
Coverage	88.99%	84.80%	82.43%
Precision 1	74.23%	75.75%	82.36%
Precision 2	76.08%	77.21%	83.42%
Precision 3	81.44%	79.96%	85.22%

Table 7: Evaluation results.

	Kafka	Goethe 1	Goethe 2
Coverage	100.00%	100.00%	100.00%
Precision 1	13.76%	9.61%	10.82%
Precision 2	13.76%	9.65%	10.86%
Precision 3	13.76%	9.69%	10.91%

Table 8.1: Baseline 1.

	Kafka	Goethe 1	Goethe 2
Coverage	100.00%	100.00%	100.00%
Precision 1	16.40%	19.30%	21.29%
Precision 2	16.40%	19.34%	21.41%
Precision 3	16.40%	19.37%	21.56%

Table 8.2: Baseline 2.

A more detailed evaluation has been performed on *Erzählungen* by Kafka. As Table 9 shows, the 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% (precision 1), 76.08% (precision 2) and 81.44% (precision 3).

Note that the cascade algorithm is "unfair" since it does not give the ending-guessing rules an opportunity to be applied unless the compound-splitting rules had failed. That is why we performed a second run with compound-splitting rules disabled and obtained much higher coverage (76.15%) and precision (66.27%, 68.43%, 74.70%). Note also, that there are some short stems, so the ending-guessing 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-guessing rules.

	Run 1			**Run 2**
	compound splitting	ending-guessing	overall (cascade)	ending-guessing only
Coverage	43.12%	45.87%	88.99%	76.15%
Precision 1	93.61%	56.00%	74.23%	66.27%
Precision 2	93.62%	57.47%	76.08%	68.43%
Precision 3	93.62%	70.00%	81.44%	74.70%

Table 9: Detailed evaluation on Kafka. The coverage is higher than in Table 6, since the "no info" column is split into SET, PART and SKIP.

7 Improvement by Linear Context

As described above, **MorphoClass** considers all successfully guessed morphological classes as equally likely. An additional module, which takes into account the immediate left noun context (defined as the two words immediately to the left) allows for a better choice between equal alternatives of morphological classes. Statistical observations are acquired from NEGRA. These are limited to the articles, prepositions and pronouns, which can be used as a left predictor of the gender, case and number of the particular noun token. For instance, according to the NEGRA corpus "*eine*" is most often followed by a feminine noun in accusative singular (see Table 10), so the **MorphoClass** hypotheses will be sorted in descending order according to the probability of the left contexts features.

eine	0.6714	Fem.Akk.Sg
	0.3213	Fem.Nom.Sg
	0.0073	Fem.Dat.Sg

Table 10: Statistics derived from NEGRA corpus.

NEGRA allowed us to acquire applicable statistics about the left context of 75% of all nouns contained in it (about 6% of the nouns had no left context of the kind we require). Our investigation of the left context rules shows that:

- the morphological class predicted by the left context rules coincides with the gender of the three most likely classes proposded by **MorphoClass** in 60% of all cases;
- the morphological class predicted by the left context rules is among the classes **MorphoClass** offers in 78% of all cases;
- The cases when a lower probability is assigned to an assumed class due to left context refinement is 14%. In this way using the linear context improves the performance in about 14% of all guesses.

8 Conclusion and Future Work

We presented some results concerning guessing morphological classes of unknown German nouns. Intuitively it is clear that 100% accuracy is impossible but the more wordforms we collect the better the guessing will be. An important feature of **MorphoClass** is that its performance can be incrementally improved by bootstrapping (remembering the new unknown wordforms belonging to the same paradigm) so the **MorphoClass**' success rate can be raised incrementally. Note that the wordforms are collected from the whole text (or from a set of texts) in a context-independent way. **MorphoClass** turns out to be a useful lexicon-acquisition tool for processing German texts.

We tried to apply the same procedure for guessing the morphological classes of unknown nouns in Bulgarian. The result is much worse (success rate less than 50%) due to the very rich inflectional morphology of Bulgarian and the impossibility to distinguish the unknown nouns in raw texts. So the relatively high precision of **MorphoClass** substantially depends on the fact that nouns can be predicted in German text with much higher certainly.

Possible directions of **MorphoClass** development are to refine the ending-guessing rules (given a much bigger lexicon), to test its performance as a component integrated in a lexical-acquisition environment for German and to apply it for other languages with relatively compact and regular morphological classes and potentially for other parts of speech.

9 Acknowledgements

We would like to thank the anonymous reviewers for the useful comments and suggestions.

References

(Adda-Decker & Adda 00) M. Adda-Decker, G. Adda. *Morphological decomposition for ASR in German.* Phonus 5, Institute of Phonetics, Saarland University, pp.129-143, 2000.

(Brill 97) E. Brill. *Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging.* In NL Processing Using Very Large Corpora. Kluwer Academic Press. 1997.

(Cucerzan & Yarowsky 00) S. Cucerzan, D. Yarowsky. *Language independent minimally supervised induction of lexical probabilities.* ACL-2000, Hong Kong, pp.270-277, 2000.

(Cutting et al. 92) D. Cutting, J. Kupiec, J. Pedersen, P. Sibun. *A practical part-of-speech tagger.* Proc. 3rd ANLP (ANLP-92), pp. 133-140, 1992.

(Daciuk 99) J. Daciuk. *Treatment of Unknown Words.* In Proc. Workshop on Implementing Automata. pp. IX-1–IX-9, 1999.

(DB/R/-MAT 92-98) *Projects DB-MAT and DBR-MAT*, see http://nats-www.informatik.uni-hamburg.de/~dbrmat/; http://www.lml.bas.bg/projects/dbr-mat/.

(DeJean 98) H. DeJean. *Morphemes as necessary concepts for structures: Discovery from untagged corpora.* In Proc. Workshop on Paradigms and Grounding in Natural Language Learning, pp. 295-299, 1998.

(DeKo 01) see http://www.ims.uni-stuttgart.de/projekte/deko/ and Säuberlich B., *Aufbau and Regelformat von DeKo*, 2001.

(Finkler 88) N. Finkler. *MORPHIX. Fast Realization of a Classification-Based Approach to Morphology.* In: Trost, H. (ed.): 4. Oster. AI-Tagung. Springer, pp. 11-19, 1988.

(Gaussier 99) E. Gaussier. *Unsuppervised learning of derivational morphology from inflectional lexicons.* In Proc. ACL Workshop Unsupervised Learning in NLP, 1999.

(Goldsmith 98) R. Goldsmith. *Automatic collection and analysis of German compounds.* Workshop on Computational Treatment of Nominals, COLING-ACL, pp. 61-69, 1998.

(Goldsmith 01) J. Goldsmith J. *Unsupervised Learning of the Morphology of a Natural Language.* Computational Linguistics, vol. 27(2), pp. 153-198, 2001.

(Hafer & Weiss 74)M. Hafer, S. Weiss. *Word segmentation by letter successor varieties.* Information Storage and Retrieval, vol. 10, pp. 371-385, 1974.

(Hietsch 84) O. Hietsch. *Productive second elements in nominal compounds: The matching of English and German.* Linguistica 24, pp. 391-414, 1984.

(Jacquemin 97) C. Jacquemin. *Guessing morphology from terms and corpora.* In Actes, 20th Ann. Int. ACM SIGIR'97, pp. 156–167, 1997.

(Kupiec 92) J. Kupiec. *Robust part-of-speech tagging using a hidden Markov model.* Computer Speech and Language, vol. 6(3), pp. 225-242, 1992.

(Lezius 00) W. Lezius. *Morphy - German Morphology, Part-of-Speech Tagging and Applications*. In Proc. 9th EURALEX Int. Congress pp. 619-623, 2000.

(Mikheev 97) A. Mikheev. *Automatic Rule Induction for Unknown Word Guessing.* In Computational Linguistics, vol. 23(3), pp. 405-423, 1997.

(Nakov et al. 02) P. Nakov, G. Angelova, W. von Hahn. *Automatic Recognition and Morphological Classification of Unknown German Nouns*. Bericht 243, FBI-HH-B-243/02, Universitaet Hamburg, 2002.

(NEGRA 01) NEGRA corpus version 2001: http://www.coli.uni-sb.de/sfb378/negra-corpus/

(Neumann 99) Neumann G., Mazzini G. (1999) *Domain-adaptive IE.* DFKI, Technical Report, 1999.

(Schmid 95) H. Schmid. *Improvements in part-of-speech tagging with an application to German*. In: Feldweg and Hinrichs, eds., Lexikon und Text, pp. 47-50, 1995.

(Schone & Jurafsky 00) P. Schone, D. Jurafsky. *Knowledge-Free Induction of Morphology Using Latent Semantic Analysis.* In Proc. of CoNLL-2000 and LLL-2000, pp. 67-72, 2000.

(Thede & Harper 97) S. Thede S., M. Harper. *Analysis of Unknown Lexical Items using Morphological and Syntactic Information with the TIMIT Corpus.* 5th Workshop on Very Large Corpora, W97-0124, 1997.

(Ulmann 95) M. Ulmann. *Decomposing German Compound Nouns.* In Proc. RANLP-95, Bulgaria, pp. 265-270, 1995.

(Van den Bosch & Daelemans 99) A. Van den Bosch, W. Daelemans. *Memory-based morphological analysis.* In. Proc. 37th Annual Meeting ACL, pp. 285-292, 1999.

(Weischedel et al. 93) R. Weischedel, R. Schwartz, J. Palmucci, M. Meteer, L. Ramshaw. *Coping with Ambiguity and Unknown Words through Probabilistic Models.* Computational Linguistics, vol. 19(2), pp. 359-382, 1993.

(Yarowsky & Wicentowski 00) D. Yarowsky, R. Wicentowski. *Minimally supervised morphological analysis by multimodal alignment.* Proc. ACL, Hong Kong, pp. 207-216, 2000.