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, Computer Science Division, USA
(2) Sofia University "St. Kl. Ohridski" and Team Vision Bulgaria
(3) Central Laboratory for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria
(4) University of Hamburg, Germany

Abstract. A system for recognition and morphological classification of unknown German nouns is described. It takes raw texts in German as input and outputs a list of the unknown nouns together with hypotheses about their stem and morphological class. The system exploits both global and local information as well as morphological properties and external linguistic knowledge. It acquires and applies Mikheev-like ending-guessing rules, which were originally proposed for POS guessing. This paper presents the system design and implementation and discusses its performance by extensive evaluation.

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 like *inflection*, *derivation* and *compounding* constantly generate new wordforms, new *proper names* appear and *foreign words* may be considered as new words as well[1]. The majority of the current systems either use lists of exceptions and proper names to support recognition of strings that look like words but do not appear in system lexicons, or apply data driven approaches to model the encountered phenomena and decide on the type and category of new words met in the text.

While the majority of the available systems for automatic processing of the unknown words aim at the recognition of the most probable part of speech (POS) tag, our system called **MorphoClass** identifies unknown wordforms in the raw text, "gathers" them into groups as candidates belonging to a single paradigm and attempts to guess the stem and morphological class of the unknown German nouns.

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 in the base paradigm. **MorphoClass** is a kind of tool for lexical acquisition: it identifies new wordforms, derives some properties and classifies unknown words from raw texts. It can be used as a tool for automatic dictionary extension with new words[2].

MorphoClass solves the "guessing" problem as a sequence of subtasks including:
• identification of unknown words (nouns only);
• recognition and grouping of the inflected forms of the same word (they must share the same stem);
• compounds splitting;
• morphological stem analysis;
• generating stem hypothesis for each group of inflected forms, and finally:
• ranking the list of hypotheses about the possible morphological class for each group of words.

This is a several-stage process, which exploits:

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

MorphoClass is not a POS guesser in its traditional meaning. The purpose of the POS guesser is to make a hypothesis about the

[1] Misspelled words or orthographic variants, resulting from e.g. the recent reform of German orthography, are also "new" but we do not consider this case here.

[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 Hamburg University and Sofia University "St. Kl. Ohridski"

possible POS for an unknown word looking at its graphemic form in the particular local context and possibly in a lexicon. **MorphoClass** is not restricted to the local context; it collects and considers all the word occurrences throughout the whole input, trying to identify other inflectional forms of the same word and derive a hypothesis for its 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 work before the tagger as well, and thus support its decisions). **MorphoClass** can be used as a lemmatiser too, as 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 as it does not conflate the inflectional and derivational forms: e.g. *generate* and *generator* would be grouped together by a IR stemmer but not by **MorphoClass**.

The paper is organized as follows. Section 2 presents some related work, which shows the large variety of existing approaches for modelling morphological phenomena in European languages. Section 3 presents **MorphoClass** resources and architecture. Section 4 discusses the ending guessing rules implemented in **MorphoClass**. The example in section 5 illustrates how **MorphoClass** works. Section 6 presents the evaluation of the **MorphoClass** performance and section 7 – its improvements by consideration of linear context. Section 8 contains the conclusion.

2. Related Work

The **MorphoClass** approach is more or less related to several classical NLP tasks, the nearest one being the morphological analysis and POS-tagging. Below we focus on the related work regarding guessing rules for POS recognition and compounds splitting in German.

German morphology. Finkler and Neumann use *n*-ary tries in their system *MORPHIX* [1]. [2] presents the *Deutsche Malaga-Morphologie* for the automatic wordform recognition for German based on *Left-Associative Grammar* using the *Malaga* system. [3] proposes general rules for morpheme boundary identification. These are hypothesised after the occurrence of sequences such as: *-ungs*, *-hafts*, *-lings*, *-tions*, *-heits*. [4] considers the problem of compound analysis by means of longest matching substrings found in the lexicon. The problem of German compounds is considered in depth by ([5], [6], [7]) and [8], who concentrates on the function of the second part of a German compound.

POS guessing. [9] uses pre-specified suffixes and performs statistical learning for POS guessing. The XEROX tagger comes with a list of built-in ending guessing rules [10]. In addition to the ending [11] exploits the capitalisation in order to guess the POS. [12] consider statistical methods for unknown words tagging using contextual information, word endings, entropy and open-class smoothing. A similar approach is presented in [13]. A very influential was the work of Brill [14], who builds more linguistically motivated rules by means of tagged corpus and a lexicon. He does not look at the affixes only but could also check their POS class in a lexicon. Mikheev proposes a similar approach that estimates the rule predictions from a raw text [15]. [16] 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 [17]. [18] proposes a *Minimum Description Length analysis* to model unsupervised learning of the morphology of European languages using corpora. [19] cuts the word, if the number of distinct letters following a pre-specified sequence surpasses some threshold, following an approach similar to [20]. [21] tries to find derivational morphology in a lexicon by means of splitting based on *p*-similarity. [22] focuses on learning morphological processes. [23] propose a memory-based approach mapping directly from letters in context to rich categories that encode morphological boundaries, syntactic class labels, and spelling changes. [24] 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 interesting work, exploiting capitalisation, as well as fixed and variable suffix is proposed in [25].

3. MorphoClass resources and architecture

Figure 1 shows the linguistic resources used in **MorphoClass** and the architecture of the main system modules.

The Stem Lexicon (SL) is compiled from resources as the NEGRA corpus and the fullform Morphy lexicon and currently contains about 13,000 German nouns (note that e.g. *der/die/das Halfter* are three lexicon items with different

morphological class each). SL facilitates the recognition of compounds, as the compound splitting module relies on noun stems from SL. The Expanded Stem Lexicon (ESL) contains all wordforms belonging to the SL entries and has been used substantially during the process of ending rules elicitation. The Word Lexicon (WL) contains important closed-class words like articles, pronouns, prepositions, which might be met in the text as part of the local context surrounding the unknown words. The inflection classes used by **MorphoClass** were designed for the DB-MAT system [26]; we reduced the original 41 classes to 39, which are not sensitive to stress alternation.

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

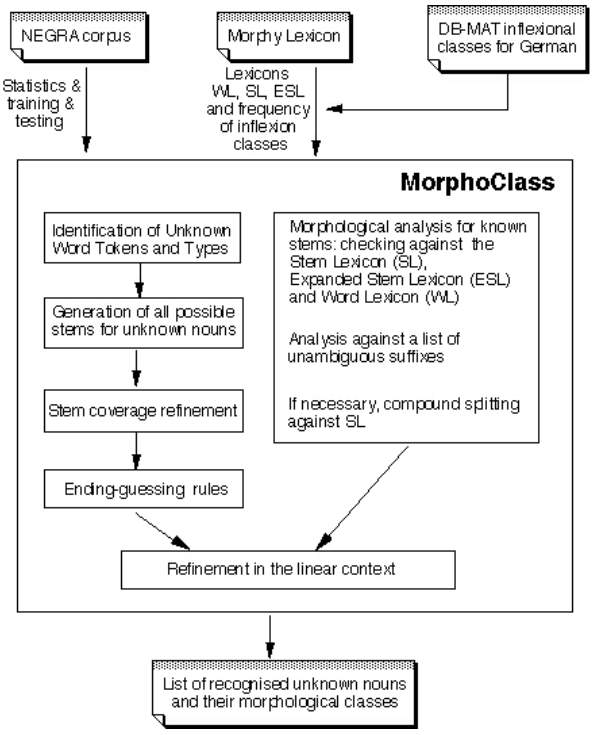

Fig 1. System Resources and Main Modules

For unknown nouns, **MorphoClass** outputs three kinds of indications:

- COMPOUND stem (successfully splitted using the available lexicon);
- ENDING RULE (an ending guessing rule has been applied);
- NO INFO (no decision was taken).

4. Ending guessing rules

We implemented Mikheev-like ending-guessing rules mechanism [15]. Mikheev originally proposed it for POS, but we adopted the approach for morphological class guessing. This resulted in 482 rules when running the rules induction against the SL and in 1789 rules when the SL entries were weighted according to their frequencies in a test corpus (see Table 1). We consider *all* endings up to 7 characters long that are met at least 10 times in the training text if after their cut at least 3 characters remain, including at least one vowel. For each noun token we extract all its endings. For each ending we collect a list of the morphological classes it appeared in, together with the corresponding frequencies. It is intuitively clear that a good ending-guessing rule is:

- *unambiguous* (predicts a particular class without or with only few exceptions),
- *frequent* (the rule must be based on large number of occurrences), and
- *long* (the longer the ending, the less is the probability that it will appear by chance, and thus the better is its prediction).

It is well-known that the *maximum likelihood estimation* (MLE) is a good predictor but it takes into account neither the rule length nor the rule frequency [15]. The *minimum confidence limit* takes into account the rule frequency but still does not prefer longer rules to shorter ones, other parameters being equal. So [15] proposes to use the logarithm of the ending length l in a score of the form:

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

$p = (x+0.5)/(n+1)$ where:

- x is the number of successful rule guesses,
- n is the total training stems compatible with the rule,
- p is a modified version of the *maximum likelihood estimation* $\hat{p}$ that ensures neither p nor $(1-p)$ could be zero;
- $\sqrt{\frac{p(1-p)}{n}}$ 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).

Mikheev in [15] scores all the rules that are met at least twice and selects only the ones above the threshold 0.65-0.80. We use a threshold of 0.90 in

order to obtain rules of higher quality (although smaller amount). For a discussions of ending-guessing rules and a full list of all rules, see [27].

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 (from lexicon)

5 Example

MorphoClass goes through the input, collects wordforms and attempts to generate one stem for each wordform group, as shown in Table 2. 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; otherwise the system tries to refine it in order to make it acceptable. 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. We are not interested 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. 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 three stems are valid since they have been obtained by reversing only legal declination rules. Stem refinement is possible after collecting more occurrences of wordforms 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 2. Largest "coverage" stems, ordered by the number of "covered" word types

How to refine Table 2 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 2. But suppose the stem *Spiel* is unknown. We have to decide that *Spiel*, *Spiele*, *Spielen* and *Spiels* are correct members of the *Spiel*-paradigm, while *Spieler* are *Spielern* are not correct and probably belong to another paradigm. 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 group - *Spieler*, *Spielern* – may be generated from *Spiel* by two classes, one masculine and one neutrum (*m3a* and *n21*). Thus, both groups are acceptable taken separately. The first group is bigger and thus it is more likely to be correct; so we decide that the first four wrodforms belong to the paradigm of *Spiel.* Applying ending-guessing rules, we will have to choose now between the four possible morphological classes (*m1*, *m9*, *n20* and *n25*). For *Spieler* and *Spielern* **MorphoClass** will continue searching another possible stem. If the two groups of wordforms 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 and the NEGRA corpus. 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. In fact it is relatively rear for a noun to have more than one morphological class: the Stem Lexicon contains only 73 such stems out of 13147 items. In our opinion, 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. We thus always look for only one paradigm for the given the stem, always preferring the biggest wordforms set that a morphological class could cover.

Table 3 is another illustration of the refinement algorithm. It lists the top unknown stems found in the NEGRA corpus ordered by the number of covered wordforms (and then alphabetically). After the refinement, the stem *Bildungsurlaube* will be deleted as a stem covering three wordforms only (see last raw of Table 3) and *Bildungsurlaub* will remain as a stem covering four wordforms (see 3rd raw of Table 3).

Unknown Stem	#	Words that Generated the Stem
Ortsbeirat	5	{ Ortsbeirat, Ortsbeirates, Ortsbeirats, Ortsbeiräte, Ortsbeiräten }
Bildungsurlaub	4	{ Bildungsurlaub, Bildungs-urlaube, Bildungsurlauben, Bildungsurlauber }
Bo	4	{ Bo, Boer, Bose, Boses }
Gemeindehaushalt	4	{ Gemeindehaushalt, Gemeinde-haushalte, Gemeindehaushaltes, Gemeindehaushalts }
Jo	4	{ Joe, Jon, Jos, Jose }
Kinderarzt	4	{ Kinderarzt, Kinderarztes, Kinderärzte, Kinderärzten }
Kunstwerk	4	{ Kunstwerk, Kunstwerke, Kunstwerken, Kunstwerks }
Lebensjahr	4	{ Lebensjahr, Lebensjahren, Lebensjahres, Lebensjahrs }
Ortsbezirk	4	{ Ortsbezirk, Ortsbezirke, Ortsbezirken, Ortsbezirks }
Stadtteil	4	{ Stadtteil, Stadtteile, Stadtteilen, Stadtteils }
Bildungsurlaube	3	{ Bildungsurlaube, Bildungs-urlauben, Bildungsurlauber }

Table 3. Unknown stems, ordered by the number of the covered wordtokens

6 Evaluation

The **MorphoClass** system has been manually evaluated over four kinds of texts:

- Reuters news, a data set of short texts containing 149 different word types, 174 word tokens;
- Franz Kafka's *Erzählungen*, 3510 word types, 13793 word tokens;
- Goethe's *Die Wahlverwandtschaften*, 10833 word types, 79485 word tokens;
- Goethe's *Wilhelm Meisters Lehrjahre*, 17252 word types, 194266 word tokens.

As we said before, **MorphoClass** considers some words as candidate-nouns (normally proper nouns and foreign words are included) and tries to decide which is the corresponding inflectional class. Sometimes the assignment is impossible (mostly when only one wordform is met in the text) and then **MorphoClass** indicates that there is not enough information of how to assign an inflectional class since neither the compound-splitting nor the ending-guessing rule were applicable. This is a positive feature of **MorphoClass,** since it avoids misleading decisions in the case of absent information. Table 4 summarises **MorphoClass** results for the four testing data sets. Note that the high percentage of "no info" on the Reuters news may be explained with the numerous foreign names in these texts. We should emphasize that **MorphoClass** always proposes a list of candidate classes but does not choose any one of them in cases of "no info".

Stems Recognised as:	Nouns	Com-pou-nds	Unknown, treated by ending-guess-ing rules:	"No info"-stems
Reuters	200	52	57 (28%)	91 (46%)
Kafka	473	185	190 (40%)	98 (21%)
Goethe 1	1706	551	837 (49%)	318 (19%)
Goethe 2	2838	896	1274 (45%)	668 (23%)

Table 4. Noun wordforms in text types

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 are covered by the ending-guessing rules. Table 4 shows that in case of longer literary texts, ending rules are applied for more than 40% of the stems, in average 45%. Their precision in isolation 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 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 was done using the 85KB text of *Erzählungen* by Kafka. We classified the stems in the following categories:

- SET - A *set* of classes is assigned by **MorphoClass** instead of a single class. About 10% of all stems were in the group of SET;
- PART - **MorphoClass** discovered a *correct* class but *not all* the correct classes in 0,6% of the cases;

- WRONG - **MorphoClass** assigned a single class and it was *wrong* (about 15%);
- YES - **MorphoClass** assigned a single class and it was the only correct one (60%);
- SKIP - The stem has been excluded from the current manual evaluation (about 10% of all stems: proper nouns, non-German nouns, non-nouns or incorrect stem).

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

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).

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).

	Run 1 compound splitting	Run 1 ending-guessing	Run 1 overall (cascade)	Run 2 ending-guesssing only
Cove-rage	43%	46%	89%	76%
Preci-sion 1	94%	56%	74%	66%
Preci-sion 2	94%	57%	76%	68%
Preci-sion 3	94%	70%	81%	75%

Table 5. Evaluation on Kafka's *Erzaehlungen*. The coverage is higher than in Table 4, since the "No info" column is split into SET, PART and SKIP

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%). Note also, that there are some short stems, so the 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.

7 Improvement by Linear Context

MorphoClass, as described in sections 3-6 above, considers all successfully guessed morphological classes as equally probable. An additional module which takes into account the left context of the nouns (article, preposition, pronoun, adjective, numerals) allows for a better choice between equal alternatives of morphological classes. The left context is defined as two consecutive words to the left of the unknown noun. The statistical observations are acquired from the NEGRA corpus. They concern all articles, prepositions, and pronouns which can be used as a left predictor of the noun gender and its case and number in the particular occurrence. For instance, from NEGRA we know that "*eine*" is most often followed by a feminine noun in accusative, singular, so the **MorphoClass** hypotheses will be sorted in descending order according to the frequency of the left contexts features.

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

Table 6. Statistics derived from NEGRA corpus.

NEGRA allowed us to acquire applicable statistics about the left context of 75% of all nouns contained there (about 6% of the nouns have no left context from the kind we use). After evaluation of the left context rules, we discovered that:

- the morphological class assumed by the "left context" rules coincides with the gender of the three most probable classes offered by **MorphoClass:** in 60% of all cases;
- the assumed class by "left context" refinement is one of the classes **MorphoClass** offers: 78%;
- Cases when lower probability is assigned to an assumed class due to left context refinement: 14%. In this way the use of linear context improves the performance in about 14% of all guesses.

8 Conclusion

In this paper we present some results concerning guessing of morphological classes of unknown German nouns. Intuitively it is clear that 100% success is impossible, but the more wordforms we collect, the better the guessing works. An important feature of our system MorphoClass is that its performance can be incrementally improved by collection of new unknown wordforms belonding to the same paradigm, so MorphoClass' success rate can be raised incrementally. Note that the wordforms are collected from the whole text (or from enlarged archive of texts) and that the wordforms are collected in a context-independent way. MorphoClass turns to be an useful lexicon-acquisition aid 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 40%) due to the very rich inflection in 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.

9 References

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

[2] Lorenz (1996). *Automatische Wortformen-erkennung fuer das Deutsche im Rahmen von Malaga.* Magisterarbeit. Friedrich-Alexander-Universität Erlangen-Nuernberg.

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

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

[5] Goldsmith, Reutter. *Automatic collection and analysis of German compounds*. *Wshop Computational Treatment of Nominals*, COLING-ACL '98, pp. 61-69.

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

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

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

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

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

[11] Weishedel et all. *Coping with ambiguity and unknown words through probabilistic models*. CL vol. 19, pp. 359-382, 1993.

[12] Thede S., Harper M. (1997) *Analysis of Unknown Lexical Items using Morphological and Syntactic Information with the TIMIT Corpus.* 5th Workshop on Very Large Corpora, August 1997.

[13] 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, Tuebingen.

[14] Brill E. (1999). *Unsupervised Learning of Disambiguation Rules for POS Tagging;* In NLP Using Very Large Corpora, 1999.

[15] Mikheev A. (1997). *Automatic Rule Induction for Unknown Word Guessing.* In CL vol 23(3), ACL 1997. pp. 405-423.

[16] Daciuk J. (1997) *Treatment of Unknown Words.*

[17] Schone, Jurafsky. *Knowledge-Free Induction of Morphology Using Latent Semantic Analysis.* In Proc. of CoNLL-2000 and LLL-2000, pp. 67-72, Lisbon.

[18] Goldsmith J. (2000) *Unsupervised Learning of the Morphology of a Natural Language.* Version of April 25, 2000. Appeared in CL (2001).

[19] DeJean H. (1998) *Morphemes as necessary concepts for structures: Discovery from untagged corpora.* In Workshop on Paradigms and Grounding in NL Learning, pp. 295-299, PaGNLL, Adelaide.

[20] Hafer M, Weiss S. (1974) *Word segmentation by letter successor varieties.* Information Storage and Retrieval, 10: 371-385, 1974.

[21] Gaussier E. *Unsuppervised learning of derivational morphology from inflectional lexicons.* ACL'99 Workshop Proceedings: Unsupervised Learning in NLP, University of Maryland, 1999.

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

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

[24] Yarowsky D. Wicentowski R. *Minimally supervised morphological analysis by multimodal alignment.* Proc. ACL-2000, Hong Kong, pp. 207-216.

[25] Cuceran S., D. Yarowsky. *Language independent minimally supervised induction of lexical probabilities*. ACL-2000, Hong Kong, pp.270-277, 2000.

[26] http://nats-www.informatik.uni-hamburg.de/~dbrmat/; http://www.lml.bas.bg/projects/dbr-mat/.

[27] Nakov, Angelova, W. v. Hahn. *Automatic Recognition and Morphological Classification of Unknown German Nouns*. Bericht 243, FBI-HH-B-243/02, Universitaet Hamburg, 2002.