A Corpus-based Approach to the Interpretation of Unknown Words with an Application to German

Stefan Klatt*

*Austrian Research Institute for Artificial Intelligence
Freyung 6/6, A-1010 Vienna, Austria
Stefan.Klatt@ofai.at

Abstract

Usually a high portion of the different word forms in a corpus receive no reading by the lexical and/or morphological analysis. These unknown words constitute a huge problem for NLP analysis tasks like POS-tagging or syntactic parsing. We present a parameterizable (in principle language-independent) corpus-based approach for the interpretation of unknown words that only needs a tokenized corpus and can be used in both offline and online applications. In combination with a few linguistic (language-dependent) rules unknown verbs, adjectives, nouns, multiword units etc. are identified. Depending on the recognized word class(es), more detailed morphosyntactic and semantic information is additionally identified in opposite to the majority of other unknown word guessing methods, which only uses a very narrow decision window to assign an unknown word its correct reading respective Part-of-Speech tag in a given text. We tested our approach by experiments with German data and received very promising results.

1. Introduction

According to (Schmid et al., 2004), usually 15-25% of the different word forms in a corpus receive no reading by the lexical and/or morphological analysis. These unknown words constitute a huge problem for NLP analysis tasks like POS-tagging or syntactic parsing.

In recent years several approaches were suggested to overcome this problem. Among them are symbolic approaches such as automatic rule induction (Mikheev, 1997) and decision tree-based methods (Orphanos and Christodoulakis, 1999) as well as statistical methods of different kinds. The approach of (Nakagawa et al., 2001) relys on support vector machines with substrings and surrounding context as relevant features. Comparing their approach on English data with the one of (Brants, 2000), they perform slightly better, which is also not surprising since the latter approach only uses a linear interpolation approach of suffixes with fixed length. (Tseng et al., 2005) propose a variety of new morphological unknown-word features based on an analytic study of POS tagging of different varieties of Mandarin Chinese. With an averaging character length of 2.4 compared to 7.7 for English, approaches like the former one seems to be inappropriate for such languages.

Nominating *the winner* of these approaches is impossible because of the following two reasons. Firstly, not all approaches were evaluated on the same language and corpus data. Secondly, some of the approaches define an unknown word (UW) in different ways. In this work, we define an UW as a word that receives no reading by the lexical and/or morphological analysis. (Nakagawa et al., 2001) and (Tseng et al., 2005) define UWs as words of the test set that do not occur in the training set. This also explains the given distribution of UWs in (Tseng et al., 2005) of their German experiment (more than 50% ordinary nouns) that totally differs of the one in our experiments (more than 50% proper nouns). Furthermore our approach differs from the former ones in that we are not interested in assigning an UW its correct reading in a text. In our opinion this could be done better by using a state-of-the-art POS tagger by offering the possible readings of which the POS tagger has to choose the *best solution*. We are mainly interested in identifying all readings of an UW as well as other of its morphosyntactic and semantic features. As a result of this we are able to compute derived word forms of UWs with regular inflection patterns and extend the lexical knowledge by these informations in a semi-automatically way.

For such tasks we consider approaches that only uses a one word respective affix window as inappropriate. To our knowledge (i) it would be impossible to interpret unknown German attributive adjectives (ADJA) with the suffixes *-e* or *-en* correctly since these suffixes are also possible German verb suffixes, and (ii) if a word has more than one reading only the one with the *most probable reading* will be assigned in all cases. But if we combine simple corpus-based information with some linguistic knowledge it will be no problem to assign the ADJA *hehren* (engl. *noble*) and the verb *wehren* (engl. *resist*) their correct readings.

The remainder of this paper is organized as follows: In the next section we describe several tests for identifying different subclasses (e.g. finite and infinite verb forms) of open word classes such as nouns (N), lexical verbs (VV) and adjectives (ADJ). In Section 3 we give a sketch how these tests are integrated into the whole system, before we evaluate our approach in Section 4. Section 5 concludes with some worthwhile extensions.

2. Architecture of Our Approach

Our approach consists of several corpus-based tests that only need a tokenized corpus. The tests are combined depending on given orthographical and/or morphological properties as well as on the so-far received results during the *application history*. Due to space limitations we can only describe the most important tests in this paper.

2.1. Word class tests

2.1.1. Tests for adjective readings

In German an adjective can occur in attributive mode as part of a noun phrase, or in predicative mode (cf. (1) and (2)).

(1) das schöne/ADJA Genua
the beautiful Genoa

(2) Genua ist schön/ADJD.
Genoa is beautiful.

In the majority of the cases an adjective occur more often as ADJA with one of the five suffix endings *-e*, *-em*, *-en*, *-er*, and *-es* than as ADJD. Furthermore an ADJA is mostly left adjacent to a noun. Since a German noun usually starts with an uppercase letter we developed the test `adjd-adja` for identifying an ADJD reading that has to fulfill the conditions that are indicated in (3). The corresponding parameters were determined by several test runs.

(3)

Parameter	Threshold
pos-ratio	50.0
neg-ratio	10.0
adjd-adja-ratio	100

For a given ADJD candidate it will be computed how often it occurs and its five possible ADJA forms and how often the ADJA forms are followed by an uppercase written word (assuming to be a noun) in the chosen corpus (cf. the values of `:adja#`, `:adjd#`, and `:adja+n-perc` in the examples).
If the condition $\frac{\texttt{:adjd\#}}{\texttt{:adja\#}} \geq$ `adjd-adja-ratio` holds the test will be immediately terminated without assigning an ADJD reading to the word, that would be the case for the adverb *schließlich* (engl. *finally*) in (4). The test will be positively terminated if an uppercase written word follows in more than `pos-ratio` % of the cases as for the candidate *hehr* (engl. *noble*) as in (5)[1]. Otherwise the test will also be neatively terminated if not more than `neg-ratio` % of the right adjacent words are uppercase written words as for the candidate *wehr* (verb stem of engl. *resist*) in (6).

```
(4)  (adjd-adja-test "schließlich")
     nil 0.0 (:lemma "schließlich")
     (:loop 1 :adja+n-perc 0.0
      :adjd# 7935 :adja# 3)

(5)  (adjd-adja-test "hehr")
     "ADJD" 1.0 (:lemma "hehr")
     (:loop 1 :adja+n-perc 94.48
      :adjd# 4 :adja# 145)

(6)  (adjd-adja-test "wehr")
     nil 0.0 (:lemma "wehr")
     (:loop 1 :adja+n-perc 0.49
      :adjd# 4 :adja# 611)
```

In German, many adjectives have also a verb reading. In such a case the former test fails, since verb readings are not very often followed by an uppercase written word. Therefore we extended the test by deleting all words that ends with a possible verb suffix (*-en* and *-e*) and apply the same procedure as before to the remaining candidate set in a second loop[2]. This enables us to assign the word *erkannt* (engl. *recognized*) an ADJD reading as in (7) as well as different verbal readings by some of our verbal tests described in Section 2.1.2.

```
(7)  (adjd-adja-test "erkannt")
     "ADJD" 1.0 (:lemma "erkannt")
     (:loop 2 :adja+n-perc 100.0
      :loop 1 :adja+n-perc 17.46
      :adjd# 1126 :adja# 504))
```

2.1.2. Tests for verb readings

In opposite to German ADJA readings that usually have a regular inflection paradigm, some German finite verbs (VVFIN) and past participle forms (VVPP) have an irregular inflection pattern, for which the development of good recognition strategies is much more difficult than for regular verb forms. In case of regular VVFIN candidates our goal is to find a derived word form in a secure context. Therefore we cut off possible verbal suffixes and try to find more than *n* occurrences (e.g. $n > 1$) of the infinitive form with a preceding infinitive marker (e.g. *zu wehren* next to a sentence end marker) in the corpus (cf. (8)). If the verb candidate starts with a known separable prefix we have to incorporate the infinitive marker into it (e.g. *einzukehren* (engl. *to stop for a bite to eat*), cf. (9)).

```
(8)  (uwi "wehren")
     (("VVFIN" 1.0
       (:lemma "wehren" :inf-suffix
        "en" :num pl :pers (1 3)
        :tempus praes ...))
      ("VVINF" 1.0
       (:lemma "wehren" ...)))

(9)  (uwi "einkehrte")
     (("VVFIN" 1.0
       (:vpfx "ein" :lemma "einkehren"
        :inf-suffix "en" :num sg
        :pers (1 3) :tempus praet...)))
```

In (8) and (9), we combined the relevant corpus-based tests in (10) and (11) by stripping off linguistically relevant suffixes as part of our unknown word interpreter (uwi) that combines the several tests as described in more detail in Section 3.

```
(10) (vvinf-test "wehren")
     "VVINF" 1.0
     (:lemma "wehren" :inf-suffix "en")
     (:zu-matches 67 :matches 573)

(11) (vvizu-test "einzukehren")
     "VVIZU" 1.0
     (:lemma "einkehren" :vpfx "ein")
     (:matches 57))
```

In the case of irregular VVFINs we consider a secure contexts as the one in which the finite verb occurs in verb second position after an uppercase written personal pronoun

[1] A result of any of our tests is a three or a four tupel with (i) the categorial result of the test respective the value `nil` if the test failed, (ii) a confidence factor (abbreviated as CF from now on) ranging from 0.0 to 1.0, (iii) further morphosyntactic information (optional), and (iv) frequency values of relevant test values.

[2] For very difficult cases we use the same strategy in a third and fourth loop to assign the words *lieb* (engl. *nice*) and *lang* (engl. *long*) their correct ADJD reading.

(assuming to be in sentence-initial position) of the set {*Ich Du Er Es Man Wir*} or before the lowercase written pronoun *er*. Since we do not find many of such configurations in a corpus, we have also defined other more insecure contexts such as lowercase-written personal pronouns to the immediate left or reflexive pronouns to the immediate right (see (12)).

(12)

Parameter	Threshold
left-sure-forms	{Ich...}
left-unsure-forms	{ich...}
right-sure-forms	{er}
right-unsure-forms	{sich...}
right-sz-forms	{. ,}
nearly-sure-ratio	7.0
sure-ratio	0.25
sz-right-ratio	30.0

Given a context window of 1 to the left and to the right, the number of adjacent words in the subsets `left-(un)sure-forms` and `right-(un)sure-forms` are assigned to the value `(un)sure-matches` as well as the numbers of `right-sz-forms` to the value `sz-matches`. The test positively terminates with CF 1.0 if one of the following two conditions is fulfilled.

(C1) $\frac{\texttt{sure-matches}*100}{|cand|} \geq$ sure-ratio

(C2) $\frac{(\texttt{sure-matches}+\texttt{unsure-matches})*100}{|cand|} \geq$ nearly-sure-ratio

In (13) this leads to the correct assignment of one irregular VVFIN form of the lemma *blasen* (engl. *blow*).

(13)
```
(vvfin-test "bläst")
"VVFIN" 1.0
(:num sg :pers (2 3)
 :modus (konj ind) :tempus praet)
(:sure-matches 9 :unsure-matches 21
 :sz-matches 30 :matches 171)
```

If the test fails, then it will be assigned a VVFIN reading with a CF of 0.6 if both conditions (C3) and (C4) are fulfilled (assuming to find a verb with a separable verb prefix in verb end position as in (14)).

(C3) The word starts with a known separable verb prefix.

(C4) $\frac{(\texttt{sz-matches}+\texttt{unsure-matches})*100}{|Wortargument|} \geq$ sz-right-ratio

(14)
```
(vvfin-test "abkamen")
"VVFIN" 0.6
(:num pl :pers (1 3)
 :modus (konj ind) :tempus praet)
(... :sz-matches 2 :matches 3)
```

The test for the identification of verb participle readings (tag VVPP) works in a similar way as the last test, since it also occurs in verb end position with a comma, sentence end marker or other typical element to its right. Due to the limited space of this paper it is impossible to explain the relevant parameter settings here in detail. Instead we present some correct interpretations and explain why the correct reading was assigned.

In (15) we stripped off the circumfix *ge-t* (as one of two possible regular past participle circumfixes in German) from the candidate, before we successfully applied the test `vvinf-test` (cf. (8)) to the rest of the string. In seven cases this word was followed by the word *worden*, the past participle form of the passive auxiliary *werden* that tells us that its left adjacent word is definitely a passivizable VVPP reading that builds the perfect with the auxiliary *have*, a very simple and elegant judgement that is superior to any statistical approaches. For the same reasons we correctly assign a VVPP reading to the word *zusammengestellt* (engl. *putting together*) in (16) by additionally stripping off the known separable verb prefix *zusammen*. In (17) we also assign the correct reading, but are not able to identify the base form because of an *orthographic irregularity*.

(15)
```
(vvpp-test "gesteigert")
"VVPP" 1.0
(:lemma "steigern" :inf-suffix "n"
 :av "haben" :passiv 7)
```

(16)
```
(vvpp-test "zusammengestellt")
"VVPP" 1.0
(:lemma "zusammenstellen"
 :vpfx "zusammen"
 :av "haben" :passiv 14)
```

(17)
```
(vvpp-test "zusammengefaßt")
"VVPP" 1.0
(:av "haben" :passiv 7)
```

2.1.3. Tests for noun readings

In the case of possible noun readings we developed tests for identifying multiword units (MWUs), locations (LOC), and to distinguish proper noun readings (NE) from ordinary noun readings (NN). In the latter case we make use of the observation that NNs are often preceded by a determiner in opposite to NEs. As a consequence we developed the test `nn-test` that counts how often a noun candidate is preceded by an determiner element.

If this is true for more than `det-left-max%` (currently 10%), then a NN reading with CF 1.0 will be assigned. If it is less than `det-left-min%` (currently 1%), then a NE reading with CF 1.0 will be assigned. If it is in between these two values, then both readings are assigned[3] (cf. the correct interpretations the person names *Berlusconi* and *Kohl*, the NN readings of *Kohl* (engl. *cabbage*) and *Salat* (engl. *salad*) in (18)-(20). An interpretation of other morphosyntactic noun features (e.g. the automatic identification of plural suffixes of NNs) is a more complex problem because of several reasons. But we are looking forward to determine a few of them (e.g. gender) by an extension of this test in the near future.

(18)
```
(nn-test "Berlusconi")
"NE" 1.0
(:det-hits 0 matches 76)
```

[3] A noun that occur 100 times in a corpus that is preceded in seven times by a determiner is assigned a NN with CF 0.7 as well as a NE with CF 0.3.

(19) `(nn-test "Kohl")`
`"NN+NE" 0.14`
`(:det-hits 65 :matches 4812)`

(20) `(nn-test "Salat")`
`"NN" 1.0`
`(:det-hits 42 :matches 148)`

NEs can be divided into several subclasses such as person, location, and organization names – as this is done for the task of named entity recognition (NER) – of which we implemented a simple corpus-based test for the recognition of locations that is based on the assumption that a location name is often preceded by a specific subset of prepositions (e.g. *in*). This test succesfully recognizes the two locations in (21) and (22).

(21) `(loc-test "Genua")`
`"LOC" 1.0`
`(:ratio 0.562 :prep-ratio 0.64`
`:prep-hits 50 :matches 98)`

(22) `(loc-test "New York")`
`"LOC" 1.0`
`(:ratio 0.701 :prep-ratio 0.222`
`:prep-hits 1335 :matches 2088)`

The latter example is different to the former ones, since for the first time a test with a multi word unit (MWU) as argument was applied. Since MWUs are a real problem for NLP tasks like parsing we developed a test that try to identify possible MWU readings. Therefore we make a bigram frequency distributions with the word and its left adjacent word as well as with its right adjacent words. If a bigram occurs more often than `sure-mwu`% (at the moment 50%) as the unigram with the lowest frequency of the two words, then we assign a MWU reading and repeat the same procedure to the next adjacent words as long as this condition holds.
This enables us to recognize the MWU *Aung San Suu Kyi* in (23). The computation of the next *best* candidate for a *MWU reading extension* was the noun *"Friedensnobelpreisträgerin* (engl. *nobel prize laureate*). Since the corresponding threshold is $<$ `sure-mwu`%, the procedure stops here. But for the task of named entity recognition this is a very important information helping us to identify *Aung San Suu Kyi* also as a female person in a very elegant way.

(23) `(mwl-test "Aung")`
`"MWU" 1.0`
`(:lemma "Aung San Suu Kyi")`
`(:next-left "Friedensnobel-`
`preisträgerin" :ratio 23.4))`

3. Combination of Word Class Tests

Figure 1 sketches the overall strategy for the interpretation of lowercase UWs. If the UW only consists of one letter, we recognize it as a letter and apply the test `fm-test` (for finding foreign material words) before we stop. If we have a word length > 3, the further analysis is triggered by the given UW suffix. In case of the suffix *-n* the submodule *n-suffix-subtest* is processed. If we assigned in this submodule the value *t* to the variable *fin* the procedure stops. Otherwise the procedure continues checking wether a VVFIN reading was assigned during the current analysis.

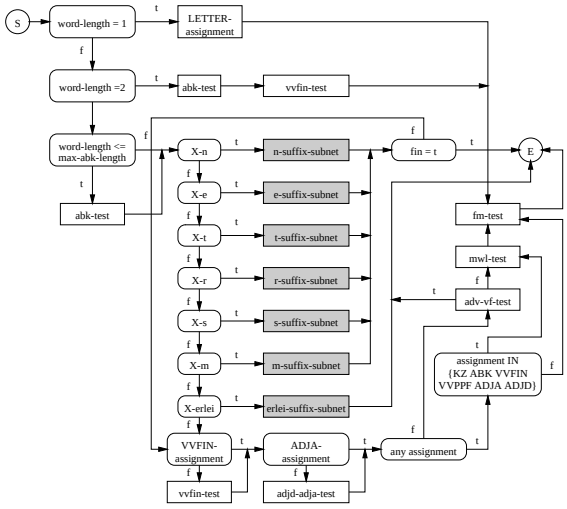

Figure 1: Processing of lowercase words

Figure 2 shows the architecture of the before mentioned submodule *n-suffix-subtest*. If the words end with one of the tree possible infinitive marker and the corresponding tests `vvinf-test` or `vvizu-test` (in the figure `vvzuinf-test`) return a VVINF or VVIZU reading, then the value *t* is assigned to the variable *fin* and the processing of this submodule stops. Otherwise some of the tests that were described in Section 2.1. are applied in the given order.

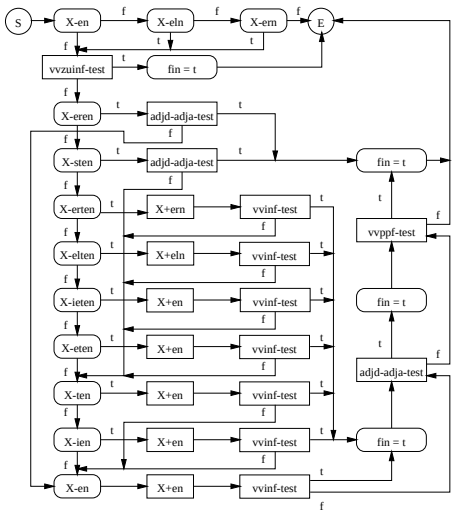

Figure 2: Processing of lowercase words with suffix *-n*

4. Evaluation

In the following we discuss the results of two experiments. Following the assumption that UWs are members of open word class tags, we extracted in the first experiment for each of such a category the first 100 different word forms of the 36 mio tokens *Stuttgarter-Zeitungs corpus* (STZ corpus) and considered these as unknown. In the second experiment, we randomly selected 500 *real* UWs of the STZ corpus received no reading by the morphology component DMOR (Schiller, 1995).

4.1. Experiment 1

Table 1 shows the result of the first experiment. The first row show the major POS tags of different ADJ, N and VV readings. For each of such a tag we selected the first 100 different word forms of the STZ corpus that were tagged by the *TreeTagger* (Schmid, 1999) accordingly (row *freq*). The row *freq-sure* show how many of these words were correctly tagged by the *TreeTagger*. The next nine rows show the number of missing, correct and spurious analyses (distinguishing analyses with a CF of 1.0 from lower ones) and the corresponding recall ($\frac{corr}{corr+miss} * 100\%$) and precision ($\frac{corr}{corr+spur} * 100\%$) rates. The last three rows show how often these words occur in the given frequency intervals in the STZ corpus.

The results for ADJA, VVFIN, VVINF, and VVPP are nearly perfect. Considering the eight missing ADJD readings, we had one irregular form, one with a major ADV reading, and one that only occur two times in our test set. Furthermore we had five words, for which we were able to assign correct ambiguous readings (VVFIN, VVPP und VVIMP). Since we know that VVPP have often an additionally ADJD reading we can increase the recall for ADJD prognoses by a corresponding extension. The reason for the missing and spurious NN and NE assignments is mainly the distribution of the left adjacent determiners. We had a few plural NN forms in the test set that were preceded by only a few determiners. On the other hand we had some NEs that are usually preceded by a determiner (country names as *die Schweiz* (engl. *Switzerland*) and river names as *der Rhein* (engl. *rhine*)). Furthermore it is difficult to assign the correct reading to words that occur less than 10 times in a corpus.

4.2. Experiment 2

Approx. 29 000 tokens of the STZ corpus are unknown to the morphology component DMOR. In this experiment we randomly selected five 100 word subsets of these UWs that occur in one of five specified frequency ranges that are indicated in the first column of Table 2. We manually annotated each of these words by its major reading (although ambiguous readings are sometimes possible). In ten cases (e.g. *Joke* as foreign material (FM) and as loan word (NN)) we assigned more than one reading as indicated in the second column. The next five columns contain the number of correct, missing, and spurious assignments and the corresponding precision and recall rates as computed as before.

As the scores show we only see significiant poorer results for the last subset that is also not surprising. Assigning every word that only occur one or two times in a corpus its correct reading is nearly an impossible task.

The hardest problem in this experiment was the distinction between NN and NE readings. Furthermore 416 words of the whole test corpus belongs to one of these readings. Therefore we evaluated these 416 words once again separately. Table 3 shows the corresponding results. The first column contains the manually annotated tags. There we find the two coarse-grained tags NN and NE, but also more fine-graided ones as locations NE(LOC). We annotated nouns that can be sometimes also assigned a FM tag by NE(FM) und NN(FM) like *Brookfields/NE(FM)* and *Hardliner/NN(FM)*, NEs with a NN base morphem by the tag NE(NN) and NEs that usually preceded by a determiner by the tag NE(DEF). NEs as part of company name were annotated as NE(CO) and acronyms as NE(ABK). Spelling errors were annotated by the tags NE(RF) and NN(RF).

	freq	corr	miss	spur	Rec	Prec
$UW_{100-500}$	100	87	13	37	87.00	70.16
UW_{51-100}	105	95	10	40	90.48	70.37
UW_{26-50}	102	92	10	38	90.20	70.77
UW_{11-25}	102	91	11	25	89.22	78.45
UW_{1-10}	101	76	25	32	75.25	70.37
$\sum$	510	438	72	146	85.88	75.00

Table 2: Evaluation of real UWs of the STZ corpus

The number of the correct NE resp. NN assignments are indicated in the first column of the five two partitioned columns, the number of wrong assignments in the second column. Looking at the subset $freq_{101-500}$, this would mean that 60 of the 70 given NE readings were correctly recognized as well as all six NN readings.

Comparing the wrong analysis in (24) with the correct one in (25), we see how we can avoid such errors in the future. Instead of only counting left adjacent determiners, it would be better to partition them into definite (def) and non-definite ones and take special care of the latter ones.

(24)
```
(nn-test "Solitude")
"NN" 1.0
(:ratio 0.28 :hits 95 :matches 337
 ("der" 80 (det def))
 ("die" 10 (det def))
 ("zur" 3 (prep-det))
 ("Die" 2 (det def)))
```

(25)
```
(nn-test "Know-how")
"NN" 1.0
(:ratio 0.33 :hits 81 :matches 242
 ("das" 38 (det def)) ...
   ... ("ihr" 12 (det possessiva))
   ... ("kein" 1 (det neg)) ...)
```

5. Conclusion and Further Work

We presented a simple parameterizable corpus-based approach that only needs a tokenized corpus for the interpretation of unknown words and demonstrated its adequacy for German data. Depending on the recognized word class other relevant morphosyntactic and semantic information is identified. As a worthwhile extension we plan to integrate more sophisticated recognition strategies to identify additional information (e.g. gender information of nouns). Furthermore we believe that this approach is also very helpful for the task of named entity recognition. Therefore we have to define tests for the recognition of person and organisation names.

Acknowledgements

This research has been sponsored by the FWF, Grant No. P16614. The OFAI is supported by the Austrian Federal Ministry of Education, Science and Culture, and the Austrian Federal Ministry for Transport, Innovation and Technology.

	ADJA	ADJD	NN	NE	VVFIN	VVINF	VVPP	VVIZU
freq	100	100	100	100	100	100	100	100
freq_{sure}	100	97	96	93	99	98	100	98
miss	1	8	6	10	2	3	2	2
$\text{corr}_{(cf=1.0)}$	99	89	79	61	74	77	98	91
$\text{spur}_{(cf=1.0)}$	–	1	6	10	–	–	1	–
$\text{corr}_{(cf<1.0)}$	–	–	11	22	23	18	–	–
$\text{spur}_{(cf<1.0)}$	–	1	1	1	–	–	–	11
Rec	99.00	91.75	93.75	89.25	97.98	96.94	98.00	91.86
Prec	100.00	97.80	92.78	88.30	100.00	100.00	98.99	89.22
$\text{Prec}_{(cf=1.0)}$	100.00	98.88	92.94	85.92	100.00	100.00	98.99	100.00
$\text{Prec}_{(cf<1.0)}$	–	0.00	91.67	95.65	100.00	100.00	–	0.00
≥ 100	87	64	74	75	77	79	82	?
10-100	11	29	12	17	18	16	15	?
< 10	2	4	10	1	4	3	3	?

Table 1: First 100 different word forms of selected POS classes from the STZ corpus

	$\text{freq}_{101-500}$		freq_{51-100}		freq_{26-50}		freq_{11-25}		freq_{1-10}		freq_{1-500}	
	NE	NN	NE	NN	NE	NN	NE	NN	NE	NN	NE	NN
NE	31	–	45	1	44	1	34	1	32	4	186	7
NE(LOC)	21	–	17	–	17	1	14	–	10	1	80	2
NE(FM)	7	–	3	–	4	1	12	4	5	4	31	9
NE(NN)	–	4	–	4	–	–	–	2	2	2	2	10
NE(DEF)	–	4	–	1	–	2	–	1	–	1	–	9
NE(CO)	1	1	–	1	–	–	2	2	–	1	3	5
NE(ABK)	–	1	–	2	1	1	–	–	–	–	1	4
NE(RF)	–	–	–	–	–	–	–	–	–	1	–	1
$\sum$ NE	60	10	65	9	66	6	63	10	49	12	303	47
	NN	NE	NN	NE	NN	NE	NN	NE	NN	NE	NN	NE
NN	3	–	9	–	10	1	11	–	10	4	43	4
NN(FM)	3	–	3	–	2	2	1	–	2	–	11	2
NN(RF)	–	–	–	–	–	–	–	–	–	–	2	3
$\sum$ NN	6	–	12	–	12	4	12	–	14	7	56	10

Table 3: Evaluation of unknown nouns

6. References

T. Brants. 2000. TnT - a Statistical Part-of-Speech Tagger. In *Proc. of the 6th Applied Natural Language Processing (ANLP-2000)*, Seattle.

S. Klatt. 2005. *Kombinierbare Textanalyseverfahren für die Korpusannotation und Informationsextraktion*. Ph.D. thesis, Universität Stuttgart.

A. Mikheev. 1997. Automatic Rule Induction for Unknown-word Guessing. *Computational Linguistics*, 23(3).

T. Nakagawa, T. Kudoh, and Y. Matsumoto. 2001. Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines. In *Proc. of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001)*, Tokyo, Japan.

G. S. Orphanos and D. N. Christodoulakis. 1999. POS Disambiguation and Unknown Word Guessing with Decision Trees. In *Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics*, Bergen, Norway.

A. Schiller. 1995. DMOR: Benutzeranleitung. Technical report, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

H. Schmid, A. Fitschen, and U. Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

H. Schmid. 1999. Improvements in part-of-speech tagging with an application to German. In S. Armstrong, K.W. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*. Kluwer, Dordrecht.

H. Tseng, D. Jurafsky, and C. Manning. 2005. Morphological features help POS tagging of unknown words across language varieties. In *Proc. of the Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju Island, Korea.