

A Morphology-System and Part-of-Speech Tagger for German^{*}

Wolfgang Lezius, Reinhard Rapp & Manfred Wettler

This paper presents an integrated tool for German morphology and statistical part-of-speech tagging which aims at making some well established methods widely available. The software is very user friendly, runs on any PC and can be downloaded as a complete package (including lexicon and documentation) from the World Wide Web. Compared with the performance of other tagging systems the tagger produces similar results.

Es wird ein integriertes Programmpaket vorgestellt, das ein Morphologie- und ein Taggingmodul für das Deutsche enthält. Die frei verfügbare Software zeichnet sich insbesondere durch hohe Benutzerfreundlichkeit aus und kann über das World Wide Web bezogen werden. Die Qualität der erzielten Ergebnisse entspricht dem derzeitigen Stand der Forschung.

1 Introduction

Morphology systems, lemmatisers and part-of-speech taggers are some of the basic tools in natural language processing. There are numerous applications, including syntax parsing, machine translation, automatic indexing and semantic clustering of words. Unfortunately, for languages other than English, such tools are rarely available, and different research groups are often forced to redevelop them over and over again. Considering German, quite a few morphology systems (Hausser 1996) and taggers (see table 1) have been developed, which are described in Wothke et al. (1993) (IBM Heidelberg), Steiner (1995) (University of Münster), Feldweg (1995) (University of Tübingen), Schmid (1995) (University of Stuttgart), Armstrong et al. (1995) (ISSCO Geneva), and Lezius (1995) (University of Paderborn). However, in most cases, the tagger is isolated from the morphology system. It relies on a lexicon of full forms which, of course, may be generated by a morphological tool. Unfortunately, most German lexicons are not available due to copyright restrictions and - as far as we know - none of them is public-domain. Therefore we have decided to make our system Morphy publicly available. It combines a morphological and tagging module in a single package and can be downloaded from the World Wide Web.¹

^{*}In: D. Gibbon, ed., *Natural Language Processing and Speech Technology. Results of the 3rd KONVENTS Conference, Bielefeld, October 1996*. Mouton de Gruyter, Berlin, 1996.

¹URL: <http://www-psycho.uni-paderborn.de/lezius/morpho.html>

Table 1: Comparison of German Taggers

	IBM Heidel- berg	Univer- sity of Münster	Univer- sity of Tübingen	Univer- sity of Stuttgart	ISSCO Geneva	Univer- sity of Paderborn
learning method	super- vised	super- vised	unsuper- vised	unsuper- vised	unsuper- vised	super- vised
context method	bi- & tri- grams	bi- grams	bi- grams	bi- grams	bi- grams	tri- grams
training corpus	20.000 words	200.000 words	200.000 words	20.000 words	20.000 words	20.000 words
test corpus	10.000 words	30.000 words	20.000 words	5.000 words	1.850 words	5.000 words
size of lexicon	534.514 words	850.000 words	500.000 words	350.000 words	30.000 words	100.000 words
tag sets large / small	689 33	143 54	- 42	- 50	- 56	456 51
accuracy large / small	77.7% 93.4%	81.5% 92.8%	- 96.7%	- 97.0%	- 96.5%	84.7% 95.9%

Since it has been created not only for linguists, but also for second language learners, it has been designed for standard PC-platforms and great effort has been put in making it as easy to use as possible.

2 The morphology module of Morphy

The morphology system is based on the Duden grammar (Drozdowski 1984). It consists of three parts: *Analysis*, *generation* and *lexical system*.

The lexical system is more sophisticated than other systems in order to allow a user-friendly extension of the lexicon. When entering a new word, the user is asked the minimal number of questions which are necessary to infer the new word's grammatical features and which any native speaker should be able to answer. In most cases only the root of the word has to be typed in, questions are answered by pressing the number of the correct alternative (see figure 1 for the dialogue when entering the verb *telefonieren*). Currently, the lexicon comprises 21.500 words (about 100.000 word forms) and is extended continuously.

Starting from the root of a word and the inflexion type as stored in the lexicon, the generation system produces all inflected forms which are shown on the screen. Among other morphological features it considers vowel mutation, shifts between *ß* and *ss* as well as pre- and infixation of markers for participles.

The analysis system for each word form determines its root and its part of speech, and, if appropriate, its gender, case, number, tense and comparative degree. It also segments compound nouns using a longest-matching rule which works from right to left. Since the system treats each word form separately, ambiguities can not be resolved. For ambiguous word forms any possible lemma

<ol style="list-style-type: none"> 1. Geben Sie den Stamm ein: telefonieren 2. Wird das Verb schwach konjugiert? <ol style="list-style-type: none"> 1: Ja 2: Nein 3. Wie lautet die 2. Person Singular Präsens? <ol style="list-style-type: none"> 1: du telefonierst 2: du telefonierest 3: du telefoniert 4. Wie lautet das Partizip des Verbs? <ol style="list-style-type: none"> 1: telefoniert 2: getelefoniert <p style="text-align: center;">Verb klassifiziert!</p>
--

Figure 1: Dialogue when entering *telefonieren* (user input is printed bold type)

and morphological description is given (for some examples see table 2). If a word form can not be recognised, its part of speech is predicted by an algorithm which makes use of statistical data on German suffix frequencies.

Morphy's lookup-mechanism when analyzing texts is not based on a lexicon of full forms. Instead, there is only a lexicon of roots together with their inflection types. When analyzing a word form, Morphy cuts off all possible suffixes, builds the possible roots, looks up these roots in the lexicon, and for each root generates all possible inflected forms. Only those roots which lead to inflected forms identical to the original word form will be selected (for details see Lezius 1994).

Naturally, this procedure is much slower than the simple lookup-mechanism in a full form lexicon.² Nevertheless, there are advantages: First, the lexicon can be kept small,³ which is an important consideration for a PC-based system intended to be widely distributed. Secondly, the processing of German compound nouns fits in this concept.

The performance of the morphology system has been tested at the Morpholympics conference 1994 in Erlangen (see Hausser (1996), pp. 13-14, and Lezius (1996)) with a specially designed test corpus which had been unknown to the participants. This corpus comprised about 7.800 word forms and consisted of different text types (two political speeches, a fragment of the Limas-corpus and a list of special word forms). Morphy recognised 89.2%, 95.9%, 86.9% and 75.8% of the word forms, respectively.

²Morphy's current analysis speed is about 50 word forms per second on a fast PC, which is sufficient for many purposes. For the processing of larger corpora we have used Morphy to generate a full-form lexicon under UNIX. This has led to an analysis speed of many thousand word forms per second.

³Only 750 KB memory is necessary for the current lexicon.

Table 2: Some examples of the morphological analysis

word form	morphological description	root
Flüssen	Substantiv Dativ Plural maskulinum	Fluß
Bauern-häusern	Substantiv Dativ Plural neutrum	Bauer / Haus
Schiffahrts-hafenmeisters	Substantiv Genitiv Singular maskulinum	Schiff / Fahrt / Hafen / Meister
Küsse	Substantiv Nominativ Plural maskulinum Substantiv Genitiv Plural maskulinum Substantiv Akkusativ Plural maskulinum Verb 1. Person Singular Präsens Verb 1. Person Singular Konjunktiv 1 Verb 3. Person Singular Konjunktiv 1 Verb Imperativ Singular	Kuß Kuß Kuß küsselfen küsselfen küsselfen küsselfen
einnahm	Verb 1. Person Singular Präteritum Verb 3. Person Singular Präteritum	(ein)nehmen (ein)nehmen
verspieltest	Verb 2. Person Singular Präteritum Verb 2. Person Singular Konjunktiv 2	ver-spielen ver-spielen
verspieltes	Adjektiv Nominativ Singular neutrum Adjektiv Akkusativ Singular neutrum	verspielt (ver-spielen) verspielt (ver-spielen)
edlem	Adjektiv Dativ Singular neutrum Adjektiv Dativ Singular maskulinum	edel edel

3 The tagging module of Morphy

Since morphological analysis operates on isolated word forms, ambiguities are not resolved. The task of the tagger is to resolve these ambiguities by taking into account contextual information. When designing a tagger, a number of decisions have to be made:

- Selection of a tag set.
- Selection of a tagging algorithm.
- Selection of a training and test corpus.

3.1 Tag Set

Like the morphology system, the tagger is based on the classification of the parts of speech from the Duden grammar. Supplementary additions have been taken from the system of Bergenholz and Schaefer (1977). The so-formed tag set includes grammatical features as gender, case and number. This results in a very complex system, comprising about 1000 different tags (see Lezius 1995). Since only 456 tags were actually used in the training corpus, the tag set was reduced to half. However, most German word forms are highly ambiguous in this system (about 5 tags per word form on average).

Although the amount of information gained by this system is very high, previous tagging algorithms with such large tag sets led to poor results in the past (see Wothke et al. 1993; Steiner 1995). This is because different grammatical features often have the same surface realization (e.g. nominative noun and accusative noun are difficult to distinguish by the tagger). By grouping together parts of speech with different grammatical features this kind of error can be significantly reduced. This is what current small tag sets implicitly do. However, one has to keep in mind that the gain of information provided by the tagger is also reduced with a smaller tag set.

Since some applications do not require detailed distinctions, we also constructed a small tag set comprising 51 tags as shown in table 3. Both tag sets are constructed in such a way that the large tag set can be directly mapped onto the small tag set.

3.2 Tagging algorithm

The tagger uses the Church-trigram-algorithm (Church 1988), which is still unsurpassed in terms of simplicity, robustness and accuracy. However, since we assumed that longer n-grams may give more information, and since we observed that some longer n-grams are rather frequent in corpora (see figure 2 for some statistics on the Brown-corpus), we decided to compare the Church algorithm with a tagging algorithm relying on variable context widths as described by Rapp (1995).

Starting from an ambiguous word form which is to be tagged, this algorithm considers the preceding word forms - which have already been tagged - and the succeeding word forms still to be tagged. For this ambiguous word form the algorithm constructs all possible tag sequences composed of the already computed tags on the left, one of the possible tags of the critical word form and possible tags on the right.

The choice of the tag for the critical word form is a function for the length of the tag sequences to the left and to the right which can be found in the training corpus. A detailed description of this algorithm is given in Rapp (Rapp 1995, pp. 149-154).

Although some authors (Cutting et al. 1992; Schmid 1995; Feldweg 1995) claim that unsupervised tagging algorithms produce superior results, we chose supervised learning. These publications pay little attention to the fact that algorithms for unsupervised tagging require great care (or even luck) when tuning some initial parameters. It frequently happens that unsupervised learning with sophisticated tag sets ends up in local minima, which can lead to poor results without any indication to the user. Such behavior seemed unacceptable for a standard tool.

Table 3: The small tag set (51 tags)

tag name	explanation of the tag	example
SUB EIG	Substantiv Eigenname	(<i>der</i>) <i>Mann</i> <i>Egon, (Herr) Hansen</i>
VER VER INF VER PA2 VER EIZ VER IMP VER AUX VER AUX INF VER AUX PA2 VER AUX IMP VER MOD VER MOD INF VER MOD PA2 VER MOD IMP	finite Verbform Infinitiv Partizip Perfekt erweiterter Infinitiv mit <i>zu</i> Imperativ finite Hilfsverbform Infinitiv Partizip Perfekt Imperativ finite Modalverbform Infinitiv Partizip Perfekt Imperativ	<i>spielst, läuft</i> <i>spielen, laufen</i> <i>gespielt, gelaufen</i> <i>abzuspielen</i> <i>lauf', laufe</i> <i>bin, hast</i> <i>haben, sein</i> <i>gahabt, gewesen</i> <i>sei, habe</i> <i>kannst, will</i> <i>können, wollen</i> <i>gekonnt, gewollt</i> <i>könne</i>
ART IND ART DEF	unbestimmter Artikel bestimmter Artikel	<i>ein, eines</i> <i>der, des</i>
ADJ ADJ ADV	Adjektivform Adjektiv, adverbiell	<i>schnelle, kleinstes</i> <i>(Er läuft) schnell.</i>
PRO DEM ATT PRO DEM PRO PRO REL ATT PRO REL PRO PRO POS ATT PRO POS PRO PRO IND ATT PRO IND PRO PRO INR ATT PRO INR PRO PRO PER PRO REF	Demonstrativpronomen, attributiv Demonstrativpronomen, pronominal Relativpronomen, attributiv Relativpronomen, pronominal Possessivpronomen, attributiv Possessivpronomen, pronominal Indefinitpronomen, attributiv Indefinitpronomen, pronominal Interrogativpronomen, attributiv Interrogativpronomen, pronominal Personalpronomen Reflexivpronomen	<i>diese (Frau)</i> <i>diese</i> <i>, dessen (Frau)</i> <i>, welcher</i> <i>mein (Buch)</i> <i>(Das ist) meiner.</i> <i>alle (Menschen)</i> <i>(Ich mag) alle.</i> <i>Welcher (Mann)?</i> <i>Wer?</i> <i>er, wir</i> <i>sich, uns</i>
ADV ADV PRO KON UNT KON NEB KON INF KON VGL KON PRI PRP	Adverb Pronominaladverb unterordnende Konjunktion nebenordnende Konjunktion Infinitivkonjunktion Vergleichskonjunktion Proportionalkonjunktion Präposition	<i>schon, manchmal</i> <i>damit, dadurch</i> <i>daß, da</i> <i>und, oder</i> <i>um (zu spielen)</i> <i>als, denn, wie</i> <i>desto, um so, je</i> <i>durch, an</i>
SKZ ZUS INJ ZAL ZAN ABK	Sonderklasse für <i>zu</i> Verbzusatz Interjektion Zahlwörter Zahlen Abkürzung	(<i>um</i>) <i>zu (spielen)</i> <i>(spielst) ab</i> <i>Wau, Oh</i> <i>eins, tausend</i> <i>100, 2</i> <i>Dr., usw.</i>
SZD SZE SZG SZK Szs SZN	Doppelpunkt Satzendezeichen Gedankenstrich Komma Semicolon sonstige Satzzeichen	:
		!?
		-
		,
		;
		()/

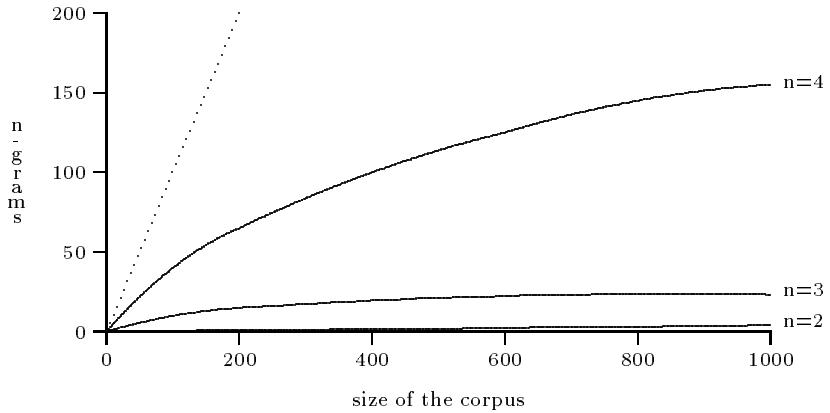


Figure 2: Statistics on the Brown corpus: number of different n-grams occurring in the corpus versus size of the corpus (all figures in thousands)

3.3 Training and test corpus

For training and testing we took a fragment from the “Frankfurter-Rundschau”-corpus,⁴ which we have been collecting since 1992. Tables and other non-textual items were removed manually. A segment of 20.000 word forms was used for training, another segment of 5.000 word forms for testing. Any word forms not recognised by the morphology system were included in the lexicon. Using a special tagging editor which - on the basis of the morphology module - for each word gives a choice of possible tags, both corpora were tagged semiautomatically with the large tag set. A recent version of the editor additionally predicts the correct tag.

4 Results

Using the probabilities from the manually annotated training corpus, the test corpus was tagged automatically. The results were compared with the previous manual annotation of the test corpus. This was done for both tagging algorithms and tag sets. For the small tag set, the Church algorithm achieved an accuracy of 95.9%, whereas with the variable-context algorithm an accuracy of 95.0% was obtained. For the large tag set the respective figures are 84.7% and 81.8%.

In comparison with other research groups, the results are similar for the small tagset and slightly better for the large tagset (see table 1). Surprisingly, inspite of considering less context, the Church algorithm performs better than the variable-context algorithm in both cases.

⁴This corpus was generously donated by the Druck- und Verlagshaus Frankfurt am Main and has been included in the CD-ROM of the European Corpus Initiative. We thank Gisela Zunker for her help with the acquisition and preparation of the corpus.



Figure 3: Accuracy versus size of the training corpus for Church’s trigram algorithm and the variable-context algorithm and both tag sets.

This is the reason why the current implementation of Morphy only includes the Church algorithm.⁵ As an example, figure 4 gives the annotation results of a few test sentences for both tag sets.

However, there are also some advantages on the side of the variable-context algorithm. First, its potential when using larger training corpora seems to be slightly higher (see figure 3). Secondly, when the algorithm is modified in such a way that sentence boundaries are not assumed to be known beforehand, the performance degrades only minimally. This means that this algorithm can actually contribute to finding sentence boundaries. And third, if there are sequences of unknown word forms in the text, the algorithm takes better guesses than the Church algorithm (examples are given in Rapp 1995, p. 155). When about 2% of the word forms in the test corpus were randomly replaced by unknown word forms, the quality of the results for the Church algorithm decreased by 0.7% for the small and by 2.0% for the large tag set. The respective figures for the variable-context algorithm are 0.9% and 1.3%, which is better overall.

In a further experiment, the contribution of the lexical probabilities to the quality of the results was examined. Without the lexical probabilities, the results decreased by 0.3% (small) and 0.6% (large tag set) for the Church algorithm, the respective figures for the variable-context algorithm were 0.9% and 0.0%.

⁵The speed of the tagger (including morphological analysis) is about 20 word forms per second for the large and 100 word forms per second for the small tag set on a fast PC.

Die	Frau	bringt	das	Essen	.
ART DEF	SUB	VER	ART DEF	SUB	SZE
Ich	meine	meine	Frau	.	
PER PRO	VER	POS ATT	SUB	SZE	
Winde	das	im	Winde	flatternde	Segel
SUB	ART DEF	PRP	SUB	ADJ	SUB
um	die	Winde			
Die	Frau	bringt			
ART DEF NOM SIN FEM	SUB NOM FEM SIN	VER 3PE SIN			
das	Essen	.	Ich		
ART DEF AKK SIN NEU	SUB AKK NEU SIN	SZE	PER NOM SIN 1PE		
meine	meine	Frau	.		
VER 1PE SIN	POS AKK SIN FEM ATT	SUB AKK FEM SIN	SZE		
Winde	das	im			
VER 3PE SIN	DEM NOM SIN NEU PRO	PRP DAT			
Winde	flatternde	Segel			
SUB DAT MAS SIN	PA1 SOL NEU AKK PLU	SUB AKK NEU PLU			
um	die	Winde	.		
PRP AKK	ART DEF AKK SIN FEM	SUB AKK FEM SIN	SZE		

Figure 4: Tagging example for both tag sets - the ambiguity rates amount to 2.4 tags per word for the small and 8.8 tags per word for the large tag set (errors are printed bold type).

5 Conclusions

We have compared two different tagging algorithms and two different tag sets. The first tagging algorithm is the Church algorithm which uses trigrams to compute contextual probabilities. The second algorithm, the so-called variable-context algorithm, has been described in paragraph 3. The smaller of the two tag sets contains 51 parts-of-speech, the larger tag set includes additional grammatical features such as case, number and gender. The small tag set is a subset of the large tag set.

In comparison with the Church algorithm, the variable-context algorithm produces similar results for the small tag set, but significantly inferior results for the large tag set. On the other hand, the performance of the variable-context algorithm for the large tag set improves faster with increasing size of the training corpus than the performance of the Church algorithm. Thus, with tagging more training texts manually, similar results are to be expected for the two algorithms.

Considering the two tag sets, the results for the small tag set are significantly better. Nevertheless, with increasing size of the training corpus an approximation of the results seems to be possible.

One of our aims for the near future is to use the output of the tagger for lemmatization. In this way a sentence like *Ich meine meine Frau.* could be unambiguously reduced to *ich / meinen / mein / Frau.*

Bibliography

- S. Armstrong, G. Russell, D. Petitpierre and G. Robert (1995). An open architecture for multilingual text processing. In: *Proceedings of the ACL SIG-DAT Workshop. From Texts to Tags: Issues in Multilingual Language Analysis*, Dublin.
- H. Bergenholz and B. Schaeder (1977). *Die Wortarten des Deutschen*. Klett, Stuttgart.
- K. Church (1988). A stochastic parts program and noun phrase parser for unrestricted text. In: *Second Conference on Applied Natural Language Processing*, pp. 136-143. Austin, Texas.
- D. Cutting, J. Kupiec, J. Pedersen and P. Sibun (1992). A practical part-of-speech tagger. In: *Proceedings of the Third Conference on Applied Language Processing*, pp. 133-140. Trento, Italy.
- G. Drosdowski (1984). *Duden. Grammatik der deutschen Gegenwartssprache*. Dudenverlag, Mannheim.
- H. Feldweg (1995). Implementation and evaluation of a German HMM for POS disambiguation. In: Feldweg and Hinrichs, eds., *Lexikon und Text*, pp. 41-46. Niemeyer, Tübingen.
- R. Hausser (1996). *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics*. Niemeyer, Tübingen.
- W. Lezius (1994). Aufbau und Funktionsweise von Morphy. Internal report. Universität-GH Paderborn, Fachbereich 2.
- W. Lezius (1995). Algorithmen zum Taggen deutscher Texte. Internal report, Universität-GH Paderborn, Fachbereich 2.
- W. Lezius (1996). Morphologiesystem MORPHY. In: R. Hausser, ed., *Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994*, pp. 25-35. Niemeyer, Tübingen.
- R. Rapp (1995). Die Berechnung von Assoziationen - Ein korpuslinguistischer Ansatz. In: Hellwig and Krause, eds., *Sprache und Computer*, vol. 16. Olms, Hildesheim.
- H. Schmid (1995). Improvements in part-of-speech tagging with an application to German. In: Feldweg and Hinrichs, eds., *Lexikon und Text*, pp. 47-50. Niemeyer, Tübingen.
- P. Steiner (1995). Anforderungen und Probleme beim Taggen deutscher Zeitungstexte. In: Feldweg and Hinrichs, eds., *Lexikon und Text*. Niemeyer, Tübingen.
- K. Woithke, I. Weck-Ulm, J. Heinecke, O. Mertineit and T. Pachunke (1993). Statistically Based Automatic Tagging of German Text Corpora with Parts-of-Speech - Some Experiments. Technical Report 75.93.02, IBM Germany, Heidelberg Scientific Center.