Morphy – German Morphology, Part-of-Speech Tagging and Applications

Wolfgang Lezius
IMS, University of Stuttgart

Abstract

In this paper we present a freely available software package that consists of a morphology tool and a part-of-speech tagger for German. The software is user-friendly, runs on any PC and can be downloaded as a complete package (including the lexicon) from the World Wide Web. One application area is searching for linguistic phenomena in annotated corpora. The corpus query tool Tatoe which is able to import Morphy's tagging output is presented as an example.

1 Introduction

Many applications in natural language processing such as syntax parsing or text indexing rely on morphology tools and part-of-speech taggers. However, these modules are often not available as a combination for the same platform. The freely available software Morphy presented here is an integrated tool which combines modules for morphology and part-of-speech tagging in a single package.[1] The user interface, which is based on the graphical environment of the Windows platform, was designed to make Morphy as user-friendly as possible.

In the first section of this paper, we outline the underlying algorithms of the morphological analysis. Afterwards, we describe the development process of the part-of-speech tagger. Finally, we discuss how linguists can profit on Morphy by using corpus query tools.

2 The morphology tool

The morphological analysis, explained in detail in [Lezius 1996a], is devided into four steps. When analysing a single word form, the first step is to check whether it might be a non-inflectional word. These words are stored in a lexicon of full forms.

In the second step, an inflectional analysis is performed for nouns, adjectives, regular and irregular verbs, and proper names. It is based on five corresponding lexicons of stems and their corresponding inflection types. The basic approach of the analysis is to determine the stem by reversing morphological processes. E.g., for the word form

Hauses all possible suffixes are isolated resulting in stem and suffix pairs (*Hauses-, Hause-s, Haus-es* etc.). Some word forms also require the reversal of vowel mutation (*Haus-Häuser, alt-älter*) or changes of *ß* and *ss* (*Fluß - Flüsse, fassen - faßt*). If there is a lexicon entry for a possible stem (here: *Haus* in the noun lexicon), the stem's inflection is generated according to the inflection type stored in the lexicon. If there is any accordance between an inflected form and the original word form, the respective morpho-syntactic description is displayed (here: genitive singular of the noun *Haus*). Thus, the approach can be characterised as analysis by generation.

Although being rather expensive in terms of processing time, the approach simplifies the treatment of compound nouns which is the following step. If the second step was unsuccessful, the word form is decomposed using a longest-matching rule. Starting from the right, the longest inflected form of a noun (cf. step two) is isolated (e.g. *Kanzleramts-ministers*). This process is reapplied on the remaining prefix (here *Kanzler-amts, Kanzler*). Some special cases like elision of a consonant are also considered (*Schif(f)-fahrt*).

If the decomposition has failed, the task of the final step is to guess the part-of-speech of the unknown word form. It relies on statistics on German suffix frequencies which have been derived by Rapp (see [RAPP 1996]). The accuracy amounts to 95%.

The morpho-syntactic categorization which is explained in more detail in [LEZIUS 1998B] among others considers gender, case, number, tense and comparative degree. Since many applications do not require such detailed distinctions, the analyser offers four degrees of formatting the output. The analysis speed amounts to 300 words per second on a standard PC. In order to illustrate the output, the analysis of a sample sentence is presented in figure 1.

Since it is part of the analyser, a generation module is also included in the morphology component. Beside the inflection of a word, all information stored in the lexicon is presented to the user (e.g. gender of a noun). The structured output may be helpful for second language learning. Since the lexicon contains 324,000 word forms (plus compound nouns) based on 50,580 stems, a wide coverage of the German language is ensured. For lexicon maintenance, a small expert system has been developed. To add a new word, first the stem has to be entered. Subsequently, the system for some characteristic forms (e.g. nominative plural of a noun) suggests inflected forms, and the user has to select the correct ones. When two up to five such questions are answered (depending on part-of-speech and inflection type) the word is classified.

3 The part-of-speech tagger

The output of the morphological analysis is often highly ambiguous. The task of the tagger is to resolve these ambiguities. Since tagging with large tag sets often has to deal with the sparse data problem, some features of Morphy's morphological categorization have been discarded (such as tense and comparative degree). The resulting tag set contains 500 tags of the original 1,500 tags. In order to develop the tagging algorithm, an adapted version of the Church tagger (see [CHURCH 1992]) was compared with an own algorithm which considers more context than the trigram-approach ([LEZIUS ET AL. 1996B]). Both algorithms were trained on a manually tagged corpus

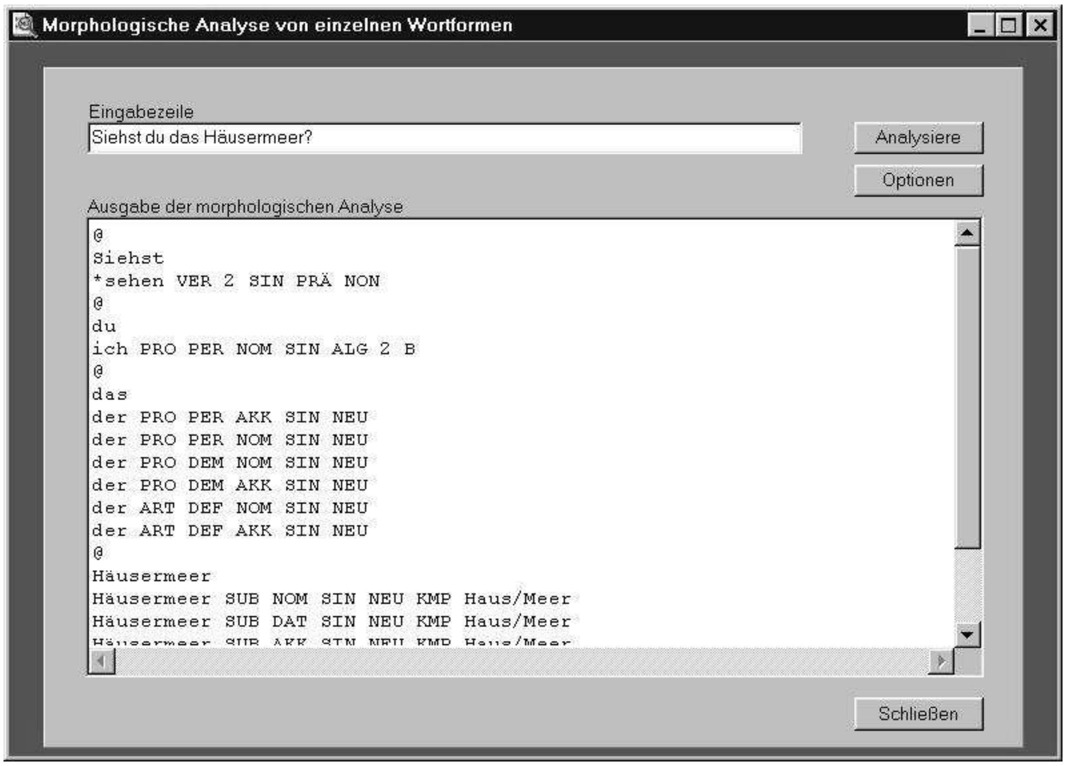

Figure 1: Morphological analysis of an example sentence

of 20,000 word forms. Tested on a small test corpus (5,000 word forms), the Church algorithm achieved better results and was therefore implemented in Morphy . The accuracy for this large tag set amounts to 85%. 60% of unknown word forms are detected correctly, for another 20% the rather slight errors are hard to avoid by the tagger (e.g. wrong gender for a noun).

However, the results are not good enough for applications requiring reliable data. Therefore also a small subset of the large tag set was constructed by excluding morpho-syntactic information. This tag set comprises 50 part-of-speech tags. Both tag sets are listed and explained in [Lezius 1998b]. The correctness for the small tag set is about 96%, for unknown word forms up to 90%. The speed of the tagger amounts to 1,000 words per second. In order to illustrate the tagger's ouput, a tagging example for the large tag set is printed in figure 2.

An obvious idea is to utilise tagging information for lemmatisation (for details see [Lezius et al. 1998a]). In the sentence *Und wie gewohnt gehe ich zur Arbeit.*, past morphological analysis it is not clear whether *gewohnt* in the present context might be an adverb or a verb, resulting in different lemmas (*gewohnt* or *wohnen*). By analysing the context, the tagger can make the correct decision (here: adverb *gewohnt*). The implemented tagger therefore also determines the most likely lemma of a word form in its respective context (cf. lemmatisation in figure 2). Tested on a text sample of 10,000 word forms, about 90% of all ambiguous word forms were lemmatised correctly.

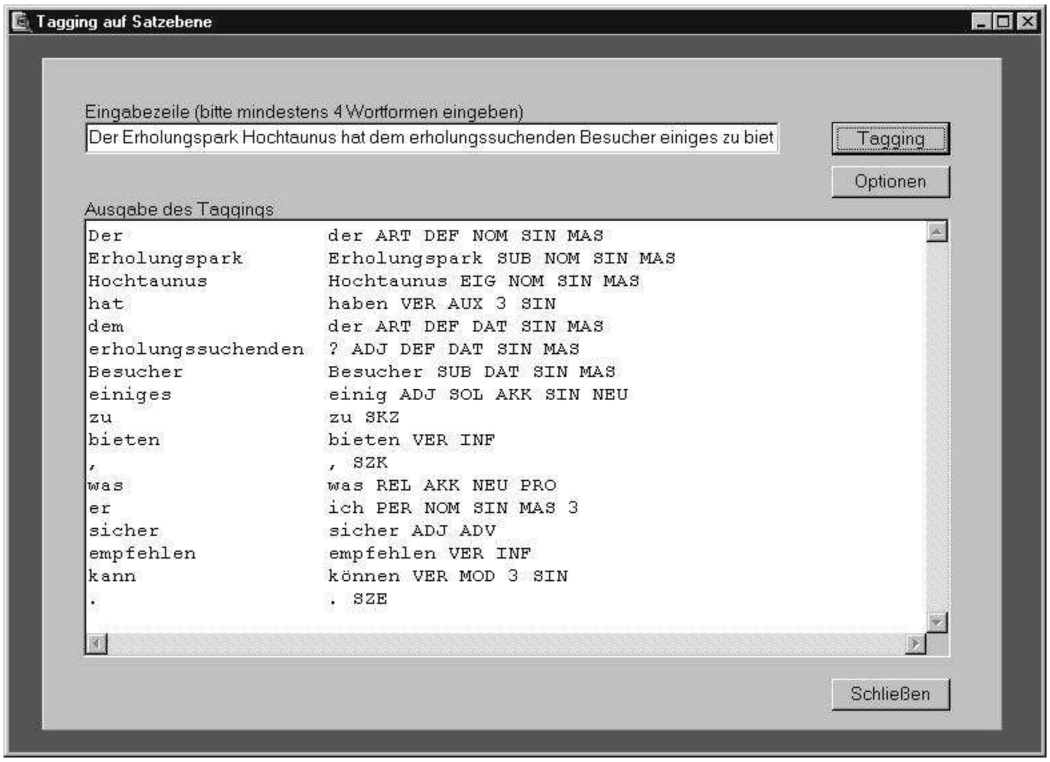

Figure 2: Tagging and lemmatisation of an example sentence

4 Searching for linguistic phenomena

Several application areas like information retrieval or syntactic analysis are based on morphological analysis and part-of-speech tagging. Another domain is the search for linguistic phenomena in a tagged corpus performed by corpus query tools.

A rather powerful but still user-friendly corpus query tool is Tatoe. It is freely available for the Windows platform[2] and imports Morphy's tagging output. A search pattern in Tatoe is a context free grammar which allows the construction of recursive rules. Assisted by a graphical environment, such queries are easy to formulate. Figure 3 shows the output for an example search pattern.

Two types of data are usually extracted by linguists: distributional data, and collocational data. These can form the basis for linguistic research, especially for lexicography. Extracting distributional data means to observe how special categories (word forms, lemmas, parts of speech) behave in different contexts. For example, one might want to find out which adverbs are used with comparatives and superlatives (pattern adverb-adjective-noun, examples: *etwas, noch, um so, weitaus, immerhin*) or as modifiers for cardinals (pattern adverb-cardinal-...-noun, examples: *beinahe, mindestens, nahezu, immerhin*). If the researcher concentrates on the question which constructions are usually used, he focusses on collocations. For example, one might extract verb-noun-collocations (pattern verb-...-noun, examples: *Respekt zollen, Rücksicht nehmen*). Another example is the extraction of idiomatic phrases like two nouns combined by a conjunction as illustrated in figure 3 (pattern noun-*und*-noun, examples: *Rat und Tat, Männer und Frauen, Lug und Betrug*).

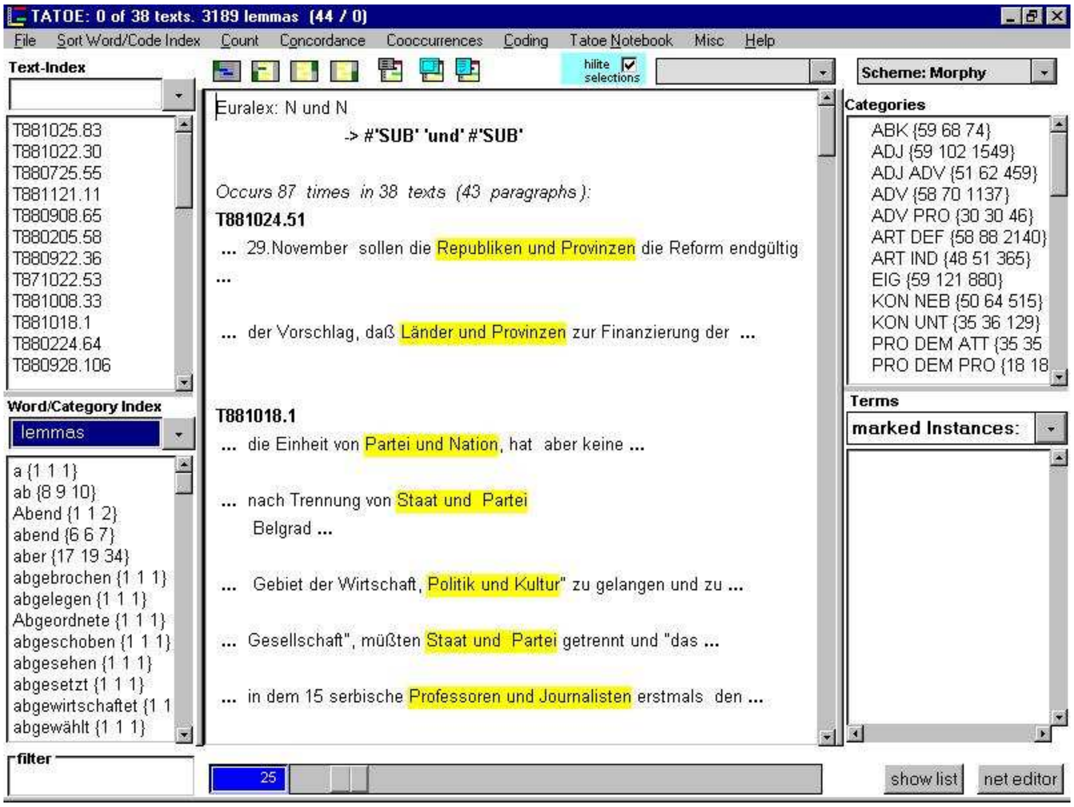

Figure 3: Occurrences matching the search pattern `SUB und SUB`

Notes

[1]Download website: `http://www-psycho.uni-paderborn.de/lezius/`

[2]Download website: `http://www.darmstadt.gmd.de/~rostek/tatoe.htm`

References

[Rostek/Alexa 1998] Lothar Rostek, Melina Alexa (1998). "Marking up in TATOE and exporting to SGML: Rule development for identifying NITF categories", in: *Computers and the Humanities*, Vol. 31/4.

[Alexa/Rostek 1996] Melina Alexa, Lothar Rostek (1996). *Computer-assisted corpus-based text analysis with TATOE.* Presented at ALLC-ACH96, Bergen, Norway. Abstracts, pp. 11-17.

[Church 1992] Kenneth W. Church (1992). *A stochastic parts program and noun phrase parser for unrestricted text.* Second Conference on Applied Natural Language Processing, Austin, Texas, pp. 136-143.

[LEZIUS 1996A] Wolfgang Lezius (1996). "Morphologiesystem Morphy", in: R. Hausser, ed., *Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994*, Niemeyer, Tbingen, pp. 25-35.

[LEZIUS ET AL. 1996B] Wolfgang Lezius, Reinhard Rapp, Manfred Wettler (1996). "A Morphology-System and Part-of-Speech Tagger for German", in: D. Gibbon, ed., *Results of the 3rd KONVENS Conference*, Mouton de Gruyter, Berlin, pp. 369-378.

[LEZIUS ET AL. 1998A] Wolfgang Lezius, Reinhard Rapp, Manfred Wettler (1998). "A freely available Morphological Analyzer, Disambiguator, and Context Sensitive Lemmatizer for German", in: *Proceedings of the COLING-ACL 1998*, pp. 743-747.

[LEZIUS 1998B] Wolfgang Lezius (1998) *Die Wortklassensysteme von Morphy* Internal Report, Universität-GH Paderborn, Fachbereich 2.

[RAPP 1996] Reinhard Rapp (1996) *Die Berechnung von Assoziationen. Ein korpuslinguistischer Ansatz.* Olms, Hildesheim.