The development of a morphosyntactic tagset for Afrikaans and its use with statistical tagging

Boris Haselbach, Ulrich Heid

University of Stuttgart, Institute for Natural Language Processing
Azenbergstraße 12, 70174 Stuttgart, Germany
{haselbbs,heid}@ims.uni-stuttgart.de

Abstract

In this paper, we present a morphosyntactic tagset for Afrikaans based on the guidelines developed by the Expert Advisory Group on Language Engineering Standards (EAGLES). We compare our slim yet expressive tagset, MAATS (Morphosyntactic AfrikAans TagSet), with an existing one which primarily focuses on a detailed morphosyntactic and semantic description of word forms. MAATS will primarily be used for the extraction of lexical data from large pos-tagged corpora. We not only focus on morphosyntactic properties but also on the processability with statistical tagging. We discuss the tagset design and motivate our classification of Afrikaans word forms, in particular we focus on the categorization of verbs and conjunctions. The complete tagset in presented and we briefly discuss each word class. In a case study with an Afrikaans newspaper corpus, we evaluate our tagset with four different statistical taggers. Despite a relatively small amount of training data, however with a large tagger lexicon, TnT-Tagger scores 97.05 % accuracy. Additionally, we present some error sources and discuss future work.

1. Introduction

The design of a part-of-speech (pos) tagset is a "trade-off between what is linguistically most desirable and computationally feasible" (Leech, 1997, p. 25). On the one hand, as much linguistic information as possible should be conveyed by annotations carried out according to a given tagset. However, on the other hand, the possibilities of statistical pos-taggers limit the number of distinctions which can be annotated with sufficient precision and thus be expressed in a tagset, as most taggers rely on differences in distribution. A tagset is thus a compromise between a precise linguistic description on the one hand, and processability on the other, i.e. the ability of the tagger to identify each class of phenomena expressed by a given tag.

In section 2.1., we present an existing tagset for Afrikaans and motivate the development of another, linguistically slim yet expressive Afrikaans tagset. We then show, in section 2.2., the criteria according to which our tagset is developed, before it is presented in section 3. Tagging results, a partial error analysis, and a comparison with the existing tagset are presented in section 4. We conclude and discuss future work in section 5.

2. Design of an Afrikaans tagset

The Expert Advisory Group on Language Engineering Standards (EAGLES) provides recommendations for the design of morphosyntactic tagsets (Leech and Wilson, 1999). To our knowledge, most morphosyntactic tagsets adhere to these recommendations. Advocating cross-linguistic common standards for tagsets and their comparability, EAGLES makes use of an attribute-value formalism which takes morphosyntactic features into account. EAGLES recommends 13 obligatory and many optional attributes. The first ones roughly correspond to the traditional classification of syntactic categories (noun, verb, adjective, adverb, etc.). The latter ones represent, for each word class, morphosyntactic subclassifications, such as, for example, type, gender, number, and case for nouns.

2.1. Motivation for the design of an Afrikaans tagset

To our knowledge, there exists only one tagset for Afrikaans, created by Suléne Pilon (Pilon, 2005). She has based her tagset on the EAGLES recommendations for the design of morphosyntactic tagsets (Leech and Wilson, 1999). Pilon's goal, among others, is to develop a linguistically expressive tagset for Afrikaans that adheres to the EAGLES standards and is processable by a pos-tagger. She develops a tagset that comprises 139 pos-tags which, on the one hand, includes morphosyntactic features proposed by EAGLES, and, on the other hand, includes semantic features, especially in the nominal and adverbial domain. For example, nouns are split into 18 morphosyntactically and semantically distinctive subgroups (and pertaining tags), e.g. measure nouns, abstract nouns, etc.

In annotation experiments with her tagset, Pilon used the TnT-Tagger (Brants, 2000), scoring an accuracy of 85.87 %. In an experiment with a condensed tagset, that only comprises the 13 major word classes proposed by EAGLES, she increased the accuracy up to 93.69 %. This result may indicate that the original tagset by Pilon is too fine-grained, especially where it encodes semantic distinctions. Statistical taggers which are trained on corpus data and thus take distributional properties of words and pos-tags into account, can often not distinguish semantic subclasses of items. Many semantically distinct lexemes have the same distribution. Consequently, the semantic distributions either must be covered by a (large) lexicon (which is hard to achieve for open word classes), or the attempt to use them with statistical taggers may involve a risk of lower tagging accuracy.

Hence, we present a new tagset which meets the conditions for statistical tagging and which, despite its reduced tagset with respect to Pilon (2005), is still morphosyntactically sufficient for a linguistic data extraction task. The morphosyntactic Afrikaans tagset, MAATS, that is presented here, is developed in the framework of research towards the extraction of Afrikaans verbal subcategorization

frames from corpora. We want to identify the subcategorization properties of verbs by means of pattern-based extraction; MAATS focuses on morphosyntactic properties of word forms.

Furthermore, MAATS is a logical tagset in the sense of Leech (1997, p. 27) which means "that the relations between the word categories symbolized by tags should be representable as a hierarchical tree [...], with attributes being inherited from one level of the tree to another." A MAATS tag is thus a sequence of letters expressing the hierarchy of EAGLES attributes which are used from left to right. For the major word classes, the letters proposed by EAGLES are used, optionally followed by letters specifying further subcategories.

2.2. Design criteria

Atwell (2008) described the criteria for tagset development as purpose-dependent. Different NLP tasks require different information from a tagset. In our case, the tagset should be both linguistically descriptive enough for the extraction of verbal subcategorization frames, on the one hand, and accurate if processed with a statistical pos-tagger, on the other.

In contrast with a highly inflecting language such as the Slavonic languages, Afrikaans shows very poor morphology. For Czech, for example, Hajič (2004) developed a positional structured tagset expressing 15 different morphosyntactic parameter. Theoretically, this tagset can accommodate over 4,200 different categorizations of word forms. From these, 216 tags were used in a corpus by Feldman and Hana (2010). For Afrikaans, however, such a fine-grained tagset is not necessary (also hardly possible to construct) due to the lack of morphological marking of morphosyntactic properties. We thus consider that a morphosyntactic tagset with much less than 100 tags is sufficient for Afrikaans.

2.2.1. Morphosyntactic focus

The primary purpose of MAATS is to support the extraction of verbal subcategorization frames. Therefore, a detailed linguistic description of some word classes is necessary. In the first place, all word classes described as obligatory by EAGLES are taken into account. Secondly, morphosyntactic features that are relevant with respect to verbal subcategorization are reflected in MAATS.

As verbs are a central element for subcategorization extraction, their distinctions are a major focus of MAATS. Afrikaans main verbs do not show much verbal inflection; however, they show morphosyntactic differences. They can be split into three different classes: simplex, particle, and prefix verbs. The former two have two distinct verb forms: (i) a base form, comprising the infinitive, all present tense finite forms, and the imperative; and (ii) a past participle formed by prefixing the verb with the morpheme *ge-*, and used, inter alia, for complex tense forms. Prefix base verbs have only one word form, as the participle is not morphologically marked by prefixing with the morpheme *ge-*. One word form represents the infinitive, finite forms, imperative, and the past participle. These properties are accounted for by the MAATS tags for main verbs which are shown in figure 1.

2.2.2. Statistical tagging focus

Statistical processability poses a problem for purely linguistically-driven tagsets, as many linguistic features are neither in complementary distribution nor translate into unambiguous morphology. The less tags a tagset comprises, the less disambiguation has to be done by a statistical pos-tagger, and with that, tagging is more accurate, but less informative.

For example, some Afrikaans adjectives inflect in attributive position and some do not (cf. adjectives in examples (1) and (2) with the noun *vrugte*, pl., "fruit").

(1) *wrang* ("bitter"): *wrang**e** vrugte*

(2) *ryp* ("ripe"): *ryp vrugte*

For Donaldson (1993, p. 163) this "has to do with the phonology of the adjective in question; [...] regardless of whether the noun is singular or plural, definite or indefinite". The rules that he gives for adjectival inflection are not explainable by the position of the adjective but only by its internal structure. Consequently, we leave such non-distributional features underspecified in MAATS.

From a strictly distributional point of view, the distinctions operated in MAATS in the verbal domain (cf. section 2.2.1.) are equally non-distributional. Therefore, we use a lexicon which provides information about the distinction between simplex, prefix, and particle verbs. It contains the base form and the past participle of approximately 6,000 simple and particle verbs, as well as 2,000 prefix verbs (cf. section 4.2.).

3. Tagset overview

Table 1 gives an overview of the tags used in MAATS.

In contrast to Pilon's tagset, nouns only have two distinct tags: common nouns vs. proper names. As this distinction might be relevant for the identification of noun phrases, it is expressed by MAATS.

As described in paragraph 2.2.1., Afrikaans verbs differ with respect to the formation of the past participle. MAATS accounts for this, and provides tags for the base form of simplex and particle verbs and their past participle, as well as one tag for prefix verbs (cf. figure 1). For auxiliary verbs, one tag for primary auxiliary verbs (*wees* "[to] be", *hê* "[to] have", and *word* "[to] be" (passive)) and one tag for modal auxiliary verbs (e.g. *kan* "can", *moet* "must", etc.) is provided, as they behave differently with respect to word order. The respective class is a closed one, so the forms are covered in the tagger lexicon.

Conjunctions trigger different subclause types and thus play an important role in the identification of verbal subcategorization frames. The coordinating conjunction *want* ("because") in example (3) triggers a clause with the verb in second position, whereas the subordinating conjunction *dat* ("that") in example (4) triggers a clause with the verb in sentence-final position.

(3) ... *want sy lees die boek.*
"... because she reads the book."

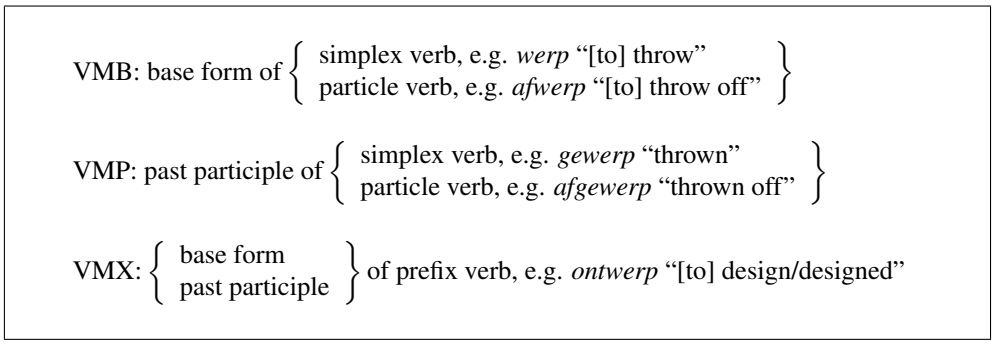

Figure 1: Verbal tags of MAATS

(4) *. . . dat sy die boek lees.*
". . . that she reads the book."

MAATS distinguishes coordinating conjunctions and subordinating conjunctions. For subordinating conjunctions, MAATS provides tags for those with a finite clause, for those with an infinitive construction, and for those with a comparative construction.
Adpositions also play an important role for verbal subcategorization as prepositions can be subcategorized by verbs. Prepositions and postpositions show distributional differences (cf. example (5) for a preposition and example (6) for a postposition). Thus, MAATS distinguishes the two types.

(5) *Jannie stap op Tafelberg.*
"Jannie is walking on Table Mountain."

(6) *Jannie stap Tafelberg toe.*
"Jannie is walking to Table Mountain."

Pronouns and determiners are relevant for the identification of noun phrases and thus they are not only strictly separated, but also their function is paid attention to, e.g. indefinite vs. personal, etc. Information about closed word classes is provided in the lexicon. However, MAATS does not distinguish between interrogative and relative pronouns/determiners. Not only do they lexically overlap (e.g. *wat* "what/that" and *wie* "who") but they also are systematically ambiguous in headless relative clauses.

(7) *Sy vra wat hy bekommer.*
"She asks what he is worried about."

(8) *Sy eet wat hy kook.*
"She eats what he cooks."

The *wat*-subclause in example (7) is subcategorized by the verb *vra* ("[to] ask") whereas in example (8) it is not subcategorized by the verb *eet* ("[to] eat"), however it is a headless relative clause. This ambiguity cannot be resolved without verbal subcategorization information (about subclauses taking verbs). This, however, is not available yet as we aim to extract exactly this information from corpora.
For adjectives, adverbs, numerals, interjections, foreign words, unclassified residuals, and punctuation marks only one tag each is provided in MAATS. First, Afrikaans does not show much inflection and/or morphosyntactic variation here, and, secondly, distinctions within these categories are not considered to be very important for verbal subcategorization.
The remaining pos-tags are due to characteristics of Afrikaans, such as the tags for particles and pronominal adverbs (e.g. the possessive particle: *Arno se nommer* "Arno's telephone number").

4. Statistical pos-tagging with MAATS

In this section, we describe the use of MAATS with different statistical part-of-speech taggers.

4.1. Taggers used

We compared four statistical taggers available under research license:

- MB-Tagger
- RFTagger
- TnT-Tagger
- TreeTagger

MB-Tagger (Daelemans et al., 1996) is a memory-based tagger storing a set of example cases extracted from a training corpus. An example case contains a word with preceding and following context on the word and part-of-speech level. Disambiguating an unknown word, MB-Tagger determines its context and uses extrapolation from the most similar cases resulting in a "best guess" of the category for the word in its context (Feldman and Hana, 2010, pp. 14–15).
RFTagger (Schmid and Laws, 2008) is a probabilistic decision tree tagger using a Markov model. It splits the pos-tags into attribute vectors and estimates the conditional probabilities of each attribute with decision trees. RFTagger is designed for fine-grained tagsets, e.g. for highly inflecting languages such as Slavonic languages, or some Germanic languages. It has been successfully used by Faaß et al. (2009) on the South African Bantu language Sepedi.
TnT-Tagger (Brants, 2000) uses a statistical trigram Markov model. Using maximum likelihood probabilities, TnT-Tagger calculates transitions and output probabilities derived from relative frequencies in the training corpus. Next to information on capitalization of words, TnT-Tagger handles unknown words by using a suffix analysis which is useful for highly inflecting languages (Feldman and Hana, 2010, pp. 8–10).

Tag	*Description*
NC	Common noun
NP	Proper name
VMB	Main verb, base form
VMP	Main verb, past participle
VMX	Main verb, prefixed
VAP	Primary auxiliary verb
VAM	Modal auxiliary verb
AJ	Adjective
AV	Adverb
AVP	Pronominal adverb
AVW	Interr./rel. pron. adverb
AT	Article
APR	Preposition
APO	Postposition
CC	Coordinating conjunction
CSI	Subordinating conj., infinitive
CSF	Subordinating conj., finite clause
CSC	Subordinating conj., comparative
PD	Demonstrative pronoun
PI	Indefinite pronoun
PS	Possessive pronoun
PW	Interr./rel. pronoun
PP	Personal pronoun
PR	Refl./reciproc. pronoun
DD	Demonstrative determiner
DI	Indefinite determiner
DS	Possessive determiner
DW	Interr./rel. determiner
NU	Numeral
IJ	Interjection
UI	Infinitive particle
UE	Existential particle
UN	Negative particle
US	Possessive particle
UV	Verbal particle
UDIS	"dis"
RF	Foreign word (residual)
RU	Unclassified residual
PU	Punctuation

Table 1: Overview of MAATS

TreeTagger (Schmid, 1994) is a probabilistic decision tree tagger using a Markov model. During training, the tagger recursively builds binary branching decision trees from trigrams. Utilizing the Viterbi algorithm, the decision trees are used to derive transition probabilities for a given state in a Markov model to determine the part of speech for an unknown word (Feldman and Hana, 2010, pp. 15–16).

4.2. Tagger lexicon

The publishing house *Pharos*[1] has provided word lists for the major word classes of Afrikaans. The lists have been semi-automatically subdivided according to the MAATS tags and manually checked. The resulting lexicon is used as the tagger lexicon. It comprises approximately 75,000 word form entries.

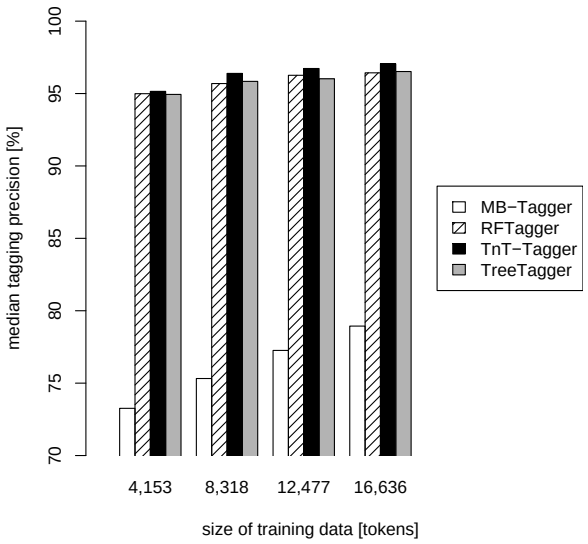

Figure 2: Median tagging results

4.3. Corpus

The corpus we used comprises text from the Afrikaans newspaper *Beeld* (years: 1989–2000; `http://www.beeld.com`; a newspaper of the *Media24* group, `http://www.media24.com`). We tokenized the corpus with Schmid's (2000) tokenizer and randomly selected a subcorpus of approximately 16,000 words. A Perl script assigned all possible MAATS tags from the above-mentioned tagger lexicon to the tokens in the subcorpus. We manually disambiguated and checked all annotations resulting in the gold standard for our experiment. [2]

4.4. Tagging evaluation

We split the corpus into four slices with an increasing number[3] of tokens: 4,153, 8,318, 12,477, and 16,636 tokens. For each slice, we calculated for each tagger the median precision using a 10-fold cross validation. This means that each tagger was ten times trained on 90 % of a given slice. The tagging result of the complementary 10 % was then compared to the gold standard.

Table 2 shows the median tagging results of MB-Tagger, RFTagger, TnT-Tagger, and TreeTagger on each slice. The results are bar-plotted in figure 2.

MB-Tagger, which does not use a lexicon for training or tagging, had the lowest precision values in our experiment. However, with increasing amounts of training data, precision of MB-Tagger clearly increased from 73.26 % on a training data set of 4,153 tokens to 78.94 % on 16,636 tokens. RFTagger, TnT-Tagger, and TreeTagger, which use a lexicon either for training or tagging, started with a

[1] We want to thank the publishing house *Pharos* making these Afrikaans word lists available to us for the present research.

[2] Thanks to Prof Elsabé Taljard and Prof Danie Prinsloo (Pretoria) for the annotation of the first 2,000 tokens, as well as to Laurette Pretorius, jr., (Pretoria) for annotating another 5,000 tokens.

[3] The numbers include sentence borders.

Size	*MB-Tagger*	*RFTagger*	*TnT-Tagger*	*TreeTagger*
4,153 tokens	73.26 %	94.99 %	95.15 %	94.94 %
8,318 tokens	75.32 %	95.69 %	96.39 %	95.84 %
12,477 tokens	77.26 %	96.26 %	96.73 %	96.02 %
16,636 tokens	78.94 %	96.43 %	97.05 %	96.52 %

Table 2: Median tagging precision values

relatively high precision (94.94 % to 95.15 %). Although the size of training data increased, the precision values of RFTagger, TnT-Tagger, and TreeTagger only slightly increased. RFTagger and TreeTagger achieved consistently more or less equal results, whereas TnT-Tagger always outperformed both of them, however not significantly. With a maximum of 16,636 tokens of training data, TnT-Tagger reached a precision of 97.05 %. This accuracy is comparable to that obtained with TnT-Tagger for other languages (e.g. for German and English TnT-Tagger scores between 96 % and 97 %; cf. Brants, 2000).

4.5. Partial error analysis

Analysing the results, we found that adjectives in predicative use were frequently tagged erroneously as adverbs (cf. examples (9) and (10)).

(9) *... sonder die nodige geriewe is alles vergeefs*$_{\text{AJ/*AV}}$.
"... without the necessary comfort, it is all in vain."

(10) *Dit is sekerlik nie maklik*$_{\text{AJ/*AV}}$ *om ...*
"This is certainly not possible that ..."

Another common source of errors are verb/noun homographs such as *druk* ("pressure/[to] press") in example (11).

(11) *Sy oefen seker te veel druk*$_{\text{NC/*VMB}}$ *op hom uit.*
"She certainly exert too much pressure on him."

Due to the large lexical overlap between pronouns and determiners, we assumed that the strict distinction of pronouns and determiners in MAATS might affect tagging precision negatively. Examples are the demonstratives *Hierdie* ("this/these") and *daardie* ("that/those") which can be both pronouns and determiners. An example for an unambiguous pronoun is *almal* ("everyone"). *Alle* ("every"), on the other side, is exclusively a determiner. However, dropping the pronoun/determiner distinction of MAATS and collapsing both categories into one, did not improve the tagging accuracy significantly. In an experiment with TnT-Tagger on 16,6336, we only scored 0.05 % better (97.10 %) without the pronoun/determiner distinction than with it. However, for modelling noun chunks for lexical data extraction we consider the pronoun/determiner distinction to be relevant. Thus, we opted for keeping this distinction in MAATS.

Comparing both Afrikaans tagsets, Pilon (2005) and MAATS, TnT-Tagger performed with MAATS (i) better than with Pilon's condensed tagset (93.69 %) and (ii) considerably better than with Pilon's original tagset (85.87 %). However, we used a much bigger tagger lexicon than Pilon (2005), which helped to score a significantly higher accuracy than with a small or with no tagger lexicon (Brants, 2000, p. 16). A similar effect was visible with MB-Tagger which does not use an external tagger lexicon but generates its own tagger lexicon from the training data. With increasing amounts of training data, and thus with an increased tagger lexicon, MB-Tagger performed considerably better than with less training data.

Nevertheless, this is an improvement, because MAATS with 39 pos-tags is more fine-grained than Pilon's condensed tagset. Indeed, it is semantically not as powerful as Pilon's original tagset. However, for a variety of NLP applications, semantic distinctions are not imperative.

5. Conclusion and future work

Pilon (2005, p. 4) concluded that 20,000 words of an Afrikaans training corpus are not enough for TnT-Tagger to compete with other state-of-the-art taggers. This may be true; however, her experiment of using only 13 pos-tags reached a significantly higher accuracy than her full tagset, which indicates that a semantically less fine-grained tagset increases accuracy. In combination with a detailed tagger lexicon, but based on training material of only 16,000 words, MAATS used with TnT-Tagger achieved an accuracy at the same level as state-of-the-art tagging studies (Feldman and Hana, 2010).

We thus conclude that MAATS in combination with a statistical tagger, especially TnT-Tagger, is well applicable to an Afrikaans corpus and that the resulting annotated corpus is usable for lexical data extraction.

For the future, we plan to analyse the tagging errors occurring in our data in greater detail in order to check for regular interferences between lexical classes their tags, and for data sparseness cases which might be solved by increasing the size of training data.

Additionally, we plan to tag the entire Afrikaans *Beeld* corpus from the years 1989 to 2000, with approximately 80 million words. This corpus will be our basis for the extraction of verbal subcategorization frames.

6. References

Eric Atwell. 2008. Development of tag sets for part-of-speech tagging. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 29.1 of *Handbooks of Linguistics and Communication Science*, chapter IV. Preprocessing corpora, pages 501–527. Mouton de Gruyter, Berlin.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA, USA.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: A memory-based part of speech tagger-generator. In Eva Ejerhed and Ido Dagan, editors, *Proceedings of the Fourth Workshop on Very Large Corpora (VLC)*, pages 14–27, Copenhagen, Denmark.

Bruce C. Donaldson. 1993. *A Grammar of Afrikaans*. Mouton de Gruyter, Berlin.

Gertrud Faaß, Ulrich Heid, Daan J. Prinsloo, and Elsabé Taljard. 2009. Part-of-speech tagging of Northern Sotho: Disambiguating polysemous function words. In *Proceedings of the EACL 2009 Workshop on Language Technologies for the African Languages (AfLaT 2009)*, pages 38–45, 31 March.

Anna Feldman and Jirka Hana. 2010. *A resource-light approach to morpho-syntactic tagging*. Number 70 in Language and Computers: Studies in Practical Linguistics. Rodopi, Amsterdam/New York.

Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum, Prague.

Geoffrey Leech and Andrew Wilson. 1999. Standards for tagsets. In Hans van Halteren, editor, *Syntactic Wordclass Tagging*, volume 9 of *Text, Speech and Language Technology*, chapter 5, pages 55–80. Kluwer Academic Publishers.

Geoffrey Leech. 1997. Introducing corpus annotation. In Roger Garside, Geoffrey Leech, and Anthony McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 1–18. Longman, London.

Suléne Pilon. 2005. Outomatiese Afrikaanse Woordsoortetikettering. Master's thesis, North-West University, Potchefstroom Campus, Potchefstroom, South Africa.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, Manchester, UK, 18–22 August.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Helmut Schmid. 2000. Unsupervised learning of period disambiguation for tokenisation. Internal Report, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.