A Probabilistic Context-free Grammar for Disambiguation in Morphological Parsing

Josée S. Heemskerk*
Institute of Language Technology and Artificial Intelligence
Tilburg University
P.O. Box 90153, 5000 LE Tilburg
The Netherlands
E-mail: joseeh@kub.nl

Abstract

One of the major problems one is faced with when decomposing words into their constituent parts is ambiguity: the generation of multiple analyses for one input word, many of which are implausible. In order to deal with ambiguity, the MORphological PArser MORPA is provided with a probabilistic context-free grammar (PCFG), i.e. it combines a "conventional" context-free morphological grammar to filter out ungrammatical segmentations with a probability-based scoring function which determines the likelihood of each successful parse. Consequently, remaining analyses can be ordered along a scale of plausibility. Test performance data will show that a PCFG yields good results in morphological parsing. MORPA is a fully implemented parser developed for use in a text-to-speech conversion system.

1 Introduction

MORPA is a MORphological PArser developed for use in the text-to-speech conversion system for Dutch, SPRAAKMAKER [van Leeuwen and te Lindert, 1993]. An important step in text-to-speech conversion is the generation of the correct phonemic representation on the basis of the input text. As is well-known, phonemic transcriptions can not be derived directly from orthographic input in Dutch, as there is no one-to-one correspondence between graphemes and phonemes. Also, stress and the effects of most phonological rules are not reflected in orthography. A text-to-speech system therefore requires an intelligent method to convert the spelled words of the input sentence into a phonemic representation.

As far as the pronunciation of words is concerned, it is impossible to list the entire vocabulary of the language, because language users have the ability to create new words and the vocabulary, as such, is indefinitely large. Daily newspapers, for instance, contain a large amount of newly formed words every day. Not all of these survive in the long run, but some of them do. Consider the examples in (1):

(1) *golfoorlog* 'gulf war'
drugsbaron 'drugs baron'
vredesmacht 'peacekeeping force'

Because it is unfeasible to give the lexicon a daily update, this approach is not appropriate if the text-to-speech system is to convert unrestricted text.

Assuming that newly created words will typically consist of already existing morphemes, and that new morphemes are added to the language only rarely, we can, however, use a lexicon in which all Dutch morphemes and their pronunciations are listed. Then complex words, such as the ones in (1), have to be decomposed into their constituent parts before their pronunciation can be looked up.

Since the pronunciation of a word does not always consists of the concatenation of the pronunciation of the morphemes, because the pronunciation of morphemes can be modified in certain contexts, the text-to-speech system also has to be provided with phonological rules which adjust the pronunciation of morphemes according to their context [Allen *et al.*, 1987;

*This work was carried out at the Phonetics Laboratory at Leiden University and supported by the Speech Technology Foundation, which is funded by the Netherlands Stimulation Project for Information Sciences, SPIN.

Nunn and van Heuven, 1993].

Dutch phonological rules are in several ways dependent on morphemic segmentation and word class assignment. As is shown in (2a), for example, the grapheme *d* is pronounced voiceless when it occurs stem-finally, but voiced when it occurs stem-initially. Final devoicing, the phonological rule which affects the pronunciation of the *d*, depends on syllable structure, and syllabification is sensitive to the morphological structure of a word: compound boundaries are also syllable boundaries. This has serious consequences in Dutch, as Dutch compounds are usually written as one word, i.e. without spaces or hyphens in between the parts. Example (2b) shows that the stress in compounds differs from the stress in monomorphemic words. In (2c) it is shown that the stress in (predicatively used) adjectival compounds differs from the stress in nominal compounds:

(2)	a	*hoofdagent*	'police sergeant'
		hoof[t] + agent	
		loofdak	'roof of foliage'
		loof + [d]ak	
	b	*avonduur*	'evening hour'
		'avond + uur	
		avontuur	'adventure'
		avon'tuur	
	c	*onecht*	'unreal'
		on + 'echt, A	
		onrecht	'injustice'
		'on + recht, N	

So to be able to produce high quality speech on unrestricted text, the text- to-speech system SPRAAKMAKER contains the morpheme lexicon-based morphological parser MORPA to recover the morphemic segmentation and word class of the input word. The module MORPHON [Nunn and van Heuven, 1993] applies phonological rules which derive the pronunciation of the word by making use of the morphological information. Also, the word class provided by MORPA feeds the module for sentence analysis which serves sentence prosody [Dirksen and Quené, 1993].

Our method of morphological analysis comprises a morpheme lexicon. Assuming that Dutch word formation is concatenative, word or word parts are recognized by dividing the word into substrings that correspond to entries in the lexicon. The major problem this method poses is ambiguity, i.e. the generation of alternative segmentations and word class assignments for one input word, many of which are implausible. In a text-to-speech system, an incorrect analysis is unacceptable, because it may lead to a wrong pronunciation [Nunn and van Heuven, 1993].

In order to deal with ambiguity, MORPA has been provided with a probabilistic context-free grammar (PCFG), i.e. it combines a "conventional" context-free morphological grammar to filter out ungrammatical segmentations with a probability-based scoring function which determines the likelihood of each successful parse. Then, aiming at a system that generates the "best" analysis first, the remaining analyses are ordered along a scale of plausibility. In this paper, I will separately describe the rule-based disambiguation techniques and probability-based scoring function. Illustrative performance data obtained from an evaluation will show that a probabilistic context-free grammar yields good results in morphological parsing.

2 Rule-based disambiguation

Decomposition of the input word is carried out in two successive stages. First, all the possible segmentations of an input word into strings of stems and affixes are generated. Secondly, each segmentation is tested for morpho-syntactic well-formedness. While the well-formedness is tested, word class is determined.

The task of recovering the morphemic segmentation with the help of a morpheme lexicon is very much complicated by the fact that a word can be segmented in more than one way. The number of alternative segmentations for an input word grows with increasing lexicon size, decreasing average length of the lexical elements and increasing average length of the input word. Our lexicon contains 17,087 entries, among which there is a large number of very small inflectional affixes. Furthermore, the input words may be very lengthy, as Dutch compounds are written as one word, and because nominal compounding, for instance, is a highly productive process. The result can be a combinatorial explosion, causing hundreds of segmentations to be generated.

In order to restrict ambiguity in the segmentation stage, we employed a number of strategies. First, we made a pragmatic operalisation of the theoretical notion "morpheme", which is traditionally defined as "the smallest meaningful unit" in word formation: in our lexicon we only listed words and affixes. Along with all simplex words and productive affixes, we listed all the word formations that belong to closed classes, i.e. words which are not formed according to productive word formation processes. Thus, our parser only has to analyse words formed according to productive rules.

Secondly, MORPA performs, if available, some tests on phonological and phonetic restrictions on the recognition of morphemes in a specific context. The ultimate effect of these tests is that incorrect recognition of highly frequent and very small inflectional suffixes, such as *-e, -t, -d, -s, -r, -n, -en* or *-er,* can be prevented in many cases.

Finally, MORPA sees to it that words belonging to minor lexical categories (such as determiners, pronouns, conjunctions, etc.) are not recognised as word parts. They never take part in morphological pro-

cesses. By rejecting these, we prevent the parser from doing work which we know beforehand will be in vain.

To illustrate the effect of the segmentation procedure, its output for the noun *beneveling* (intoxication) is shown in (3)[1]:

(3) a be + neef + eling
b be + neef + e + ling
c be + nevel + ing
d been + e + veel + ing
e be + n + e + veel + ing
f be + neef + eel + ing

All of the parts in the segmentations under (3) are Dutch morphemes listed in the morpheme lexicon. Because the segmentation procedure analyses the input word into all possible strings of morphemes without any further grammatical knowledge, it generates along with the one and only plausible segmentation be + nevel + ing (3c), several alternative segmentations. Many of these violate grammatical and/or semantic restrictions.

In order to filter out ungrammatical segmentations, each segmentation is checked for its morphosyntactic well-formedness with the help of a categorial grammar. Consequently, every segmentation that is not in accordance with the rules of Dutch morphology is rejected by the parser. While checking, the word class of the grammatical segmentations is determined.

In accordance with the principles of Categorial Grammar, our parser does not make use of a set of explicitly represented word formation rules. Instead, the morphological subcategorisation information is encoded in the form of category assignments in the lexicon. That is, prefixes have been assigned a category of type A/B, which means that they take a stem of category A on their right-hand side to yield a word of category B[2]. For instance, the prefix *be-* with category N/V requires a nominal stem to the right to form a verb. Likewise, suffixes of category $A\backslash B$ look for a stem of category A on their left-hand side to yield a word of category B. Thus, the suffix *-ing*, $V\backslash N$, requires a verbal stem to the left to form a noun. Free morphemes, such as *nevel*, are assigned primitive categories, such as V or N[3].

In a strictly bottom-up fashion, the parser iteratively attempts to combine two adjacent elements, reducing them in accordance with their categorial specification with the help of three very general reduction laws:

(4) prefixation: $A/B \cdot A \rightarrow B$
suffixation: $A \cdot A\backslash B \rightarrow B$
compounding: $A \cdot B \rightarrow B$

For pragmatic reasons, MORPA's rule for compounding is not a categorial rule, but a categorial-like rule: two adjacent stems AB may, according to the Right-Hand Head Rule be combined into a word of category B[4]. In addition to this general rule for compounding, the grammar contains a small set of rules defining productive compounding. An analysis fails as soon as a string of categories cannot be reduced to one single category.

The examples in (5) illustrate how iterative categorial reduction results in a successful parse. The structures show the derivation and determination of the output category of (3c). Also, the examples in (5) illustrate that, while the categorial grammar filters out many ungrammatical segmentations and derives the word class of the input word, parsing introduces a new type of ambiguity: one segmentation can be assigned more than one structure. The ambiguity in (5) is due to the fact that the morphemes *be-* (en-) and *nevel* (mist) can belong to more than one lexical category and as a consequence can be reduced in more than one way. The ambiguity in (5a) and (5b), is spurious in the sense that it does not correlate with a difference in pronunciation or word class assignment. The reduction in (5c) results in an incorrect word class assignment.

Because the word syntax as such is not restrictive enough, it was supplemented with a component which heavily restrains the parser in building structures. This component, which is inspired by Lexical Phonology, imposes an ordering on the attachment of affixes and stems. Consequently, it restricts the type or the complexity of the stem that an affix or other stem may attach to. Rejection of structures can result in avoiding incorrect word class assignment and rejection of incorrect segmentations.

In Lexical Phonology, the interaction between stress behaviour and affix order is explained. [Chomsky and Halle, 1968] distinguished two classes of suffixes with different stress properties, and [Siegel, 1979] observed that this distinction correlates with the order in which the suffixes attach. Over the years, theoretical linguists have become sceptical

[1] When segmenting, MORPA takes into account that Dutch word stems, when inflected or used as the base of a derivation, may undergo spelling changes. It would take us too far to go into the spelling rules here, but in (3) the effect of rules such as 'vowel gemination' and 'devoicing of stem-final consonants' shows up. See for more detail [Heemskerk and van Heuven, 1993].

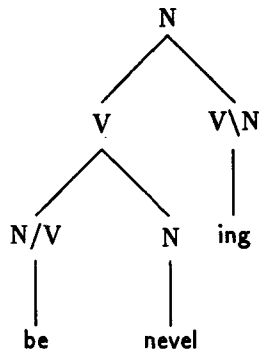
[2] Note that in the literature on categorial grammar the notational variant B/A is frequently used.

[3] Since our parser only accounts for morphological subcategorisation, the set of lexical categories does not equal the set of syntactic categories. For example, all verbs are assigned category V, irrespective of (in)transitivity. The use of syntactic categories would complicate the grammar considerably. See [Dowty, 1979] and [Moortgat, 1987] for a discussion on this matter.

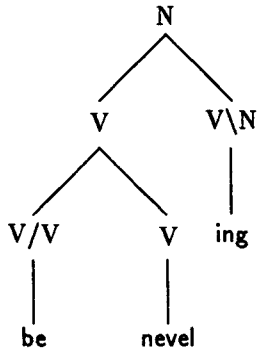
[4] For more principled approaches see [Hoeksema, 1984; Moortgat, 1987]

of these "level theories", because of the so-called "bracketing paradoxes", i.e. constructions in which two distinct constituent structures (for instance a morphological and a phonological one) have to be assigned to a word[5]. Despite the occurrence of bracketing paradoxes, however, the claims on level ordered morphology following from these theories are highly interesting: in checking the morphological claims which follow from one of the theories that have been developed for Dutch, [van Beurden, 1987], against a large database containing approximately 123,000 Dutch words, relatively few counter-examples were found.

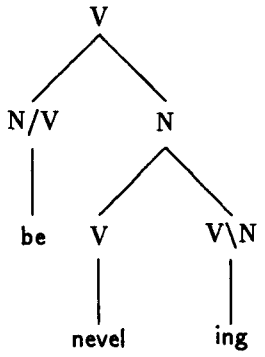
(5) a



b



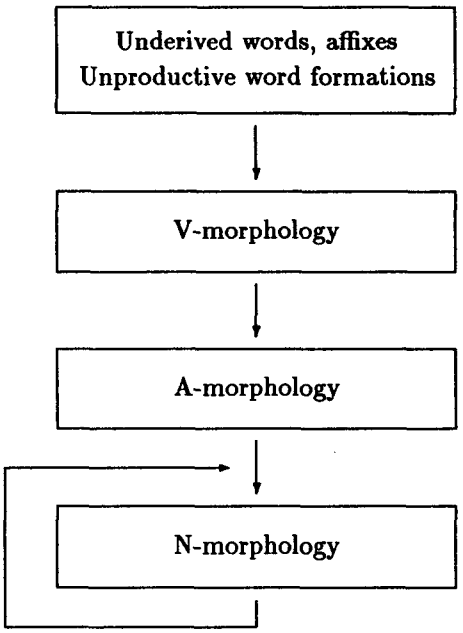
c



[5]See for a recent discussion of this topic [Spencer, 1991]

Van Beurden claims that affix order does not depend on stress properties, but on categorial properties. Thus, the major characteristic of this model is that each attachment level is associated with a specific lexical output category. The model seems particularly suitable for use in MORPA, because it is easy to integrate with our categorial parser. The model implemented in MORPA, shown in (6), is an extension of Van Beurden's model in a way which is consistent with its basic assumptions[6].

(6)



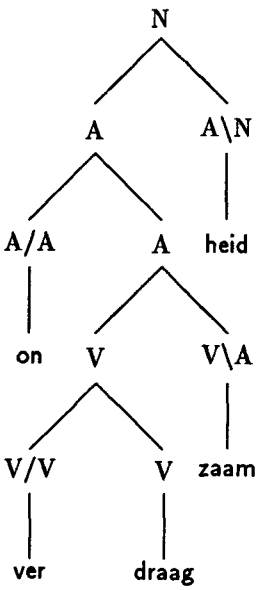
On the basis of this model, the Dutch vocabulary can be divided into four levels. Each of the levels in (6) may be viewed as possible successive stages in word formation. The first level, or lexical level, comprises the lexicon of simplex words, affixes and irregular formations. This level also contains all (borrowed) Romance words. The elements of this lexical level may be successively developed on the second level on which V(erbal)-morphology takes place; the third level on which A(djectival)-morphology takes place and the fourth level on which N(ominal)-morphology takes place. The name of the level indicates the resulting word class. Each of these levels preserves the possibility for suffixation, compounding and prefixation. On the levels for V-morphology and A-morphology each of these processes may take place

[6]In van Beurden's model each categorial level has a phonological level associated with it. As we are mainly interested in the morphological aspects, we leave the phonological claims for what they are: within SPRAAK-MAKER, MORPA and MORPHON (the phonological module) are autonomous modules, and as MORPA precedes MORPHON, any interaction between the two systems is one way.

only once. We assume that only the processes on the N-morphology level are recursive, i.e. may take place more than once (see [Heemskerk, 1989] for more details).

The model correctly predicts the derivation of the word *onverdraagzaamheid* (intolerance). As shown in (7), first verbal prefixation yields the verbal stem *verdraag* (tolerate), then adjectival suffixation yields the adjective *verdraagzaam* (tolerant), adjectival prefixation yields the adjective *onverdraagzaam* (intolerant) and, finally, nominal suffixation yields the noun *onverdraagzaamheid* (intolerance):

(7)



Also, the level module rules out the analysis in (5c): the nominal suffix *-ing* must not be attached before the verbal prefix *be-*. Therefore the word cannot be analysed as a verb.

(8)

Segmentations	word class assigned by	
	categ. grammar	level module
be + neef + eling	N	N
be + neef + e + ling	N	-
be + nevel + ing	N V	N -
be + neef + eel + ing	N	-

If we return to the example of *beneveling* we find that of the six alternative segmentations in (3), only four are accepted by the categorial component. As is shown in (8) one of these segmentations has been assigned a wrong word class. In (8) it is also shown that, as a result of the level ordering, three of the assigned word classes (and matching structures[7]) were rejected. Consequently, two analyses remain.

3 Probability-based scoring function

Clearly, the ultimate handling of the remaining ambiguity in (8) demands recourse to semantics and world knowledge. For the large-scale domain we are dealing with, however, we considered it unfeasible to implement semantic and pragmatic constraints. Thanks to the availability of a large annotated corpus, the alternative of constructing a PCFG came within reach. The corpus, being a representative sample of the past or existing vocabulary, is expected to capture implicitly various semantic and pragmatic constraints. [Fujisaki *et al.*, 1989; Liberman, 1991]. Empirical estimation of the probability of a parse tree on the basis of the corpus enables us to order the competing analyses along a scale of plausibility and select the "best" parse out of the set of alternatives.

A parse tree, such as (5a), is a series of applied production rules[8]. In a context-free grammar it is assumed that the application of a production rule is independent of previously applied rules. In a PCFG, each production rule r is assigned an estimated probability of use and the probability of the parse tree t is the product of the constituting production rules $r_1, r_2, \ldots, r_m$:

$$(9) \quad P(t) = P(r_1) \times P(r_2) \times \ldots \times P(r_m)$$

The probability of each production rule in the grammar has been estimated by means of straightforward counting of appearances in the corpus, resulting in relative frequencies. Let G be any non-terminal symbol of the grammar; $n(G)$ the number of productions rewriting G and $P(i|G)$ the probability that the ith of these productions takes place, then

$$(10) \quad P(i|G) = \frac{n(i|G)}{n(G)}$$

It is assumed that for all $i = 1, 2, \ldots, n(G)$, $P(i|G)$ is a positive number and that $\sum_i P(i|G) = 1$.

[7] In (8), I abstract from hierarchical structures, since they are irrelevant for pronunciation. Relevant for pronunciation are the morphemic segmentation and word class assignment. Consequently, the structures of (5) are represented as the segmentation be + nevel + ing, which has been assigned two word classes N and V.

[8] In this section, I will give a top-down description of a parse tree and discuss production rules of the type "$A \rightarrow B\ C$", rather than bottom-up reduction and rules of the sort "$B\ C \rightarrow A$" used by the parser.

MORPA's grammar comprises three different types of production rules:

(11) a $w \rightarrow T$
b $T \rightarrow N1\ N2$
c $N \rightarrow M$

In (11) w is the start symbol for words[9], T any member of the set of atomic categories which are possible top nodes: $\mathcal{T} = \{n, v, a, \ldots\}$, N any member of the set of non-terminals containing both atomic and functor categories: $\mathcal{N} = \{n, n/v, v\backslash n, v, \ldots\}$, $\mathcal{T} \subset \mathcal{N}$, and M any member of the set of terminals: $\mathcal{M} = \{\textsf{be}, \textsf{nevel}, \textsf{ing}, \ldots\}$.

The probability of (5a) is then determined as in (12)[10]:

(12) $P([n\ [v\ [n/v\ \textsf{be}][v\ \textsf{nevel}]][v\backslash n\ \textsf{ing}]]) =$
$P(w \rightarrow n) \times$
$P(n \rightarrow v\ \ v\backslash n) \times$
$P(v \rightarrow n/v\ \ n) \times$
$P(n/v \rightarrow \textsf{be}) \times$
$P(n \rightarrow \textsf{nevel}) \times$
$P(v\backslash n \rightarrow \textsf{ing})$

Thus, this simple PCFG provides general information on how likely a parse tree is going to appear.

It is well-known that the accuracy of the empirical estimate of a probability function depends heavily on the appropriateness of the training set: for one thing, it must have a reasonable size and be representative of the domain that is being modelled. Our training set was the CELEX database which contains approximately 123,000 Dutch stems provided with syntactic information, a morphological decomposition and token frequency information [van der Wouden, 1988; Burnage, 1990]. The token frequency information is based on a 44-million-word corpus. We collected from this database both type and token frequencies: type frequencies indicate how often a production rule occurs in the Dutch vocabulary (i.e. in the 123,000 stems corpus); token frequencies indicate how often a production rule occurs in Dutch texts (i.e. in the 44-million-word corpus). The underlying idea was that for tests on dictionary samples the empirical estimate must be based on type frequencies, whereas for tests on text samples it must be based on token frequencies.

Given the information in the database, we expected the collection of frequency data to be a matter of straightforward counting: CELEX's morphological decomposition consists of hierarchical structures which are comparable to MORPA's structures (*cf.* the examples in (5)), the syntactic information consists of the word class, and because each stem in the stem corpus is provided with a token frequency, type and token frequencies could be collected simultaneously: every time a production rule was encountered in the stems corpus, 1 was added to its type frequency, and the token frequency of the word in which the rule was attested was added to its token frequency.

Unfortunately, however, straightforward counting of all production rules contained in CELEX did not suffice to provide MORPA with the relevant information: it turned out that the set of production rules employed by MORPA was not contained in the set of production rules given by CELEX. For a very large part, the mismatch between the rules is caused by the fact that CELEX and MORPA yield different analyses. For example, because in MORPA all words formed according to unproductive rules are entirely listed in the lexicon, and the Dutch adjectival suffix *-elijk* '-ly' is considered to be unproductive, all words derived by this suffix are listed. In CELEX, however, these words are decomposed. Now, in order to analyse the word *vriendelijk* (friendly), MORPA will employ the production rule (13a), whereas CELEX employed the rules in (13b):

(13) a $A \rightarrow \textsf{vriendelijk}$
b $A \rightarrow N\ \ N\backslash A$
$N \rightarrow \textsf{vriend}$
$N\backslash A \rightarrow \textsf{elijk}$

Consequently, straightforward counting of the production rules in CELEX, would result in overestimating the probability of the productions "$A \rightarrow N\ N\backslash A$" and "$N \rightarrow$ vriend", and lack of frequency information for the production "$A \rightarrow$ vriendelijk".

Amongst the MORPA rules which were not contained in the set of CELEX rules, there were also all the rules introducing inflectional affixes and inflected stems. Of course, this is due to the fact that the 123,000-entry corpus only contains stems. As CELEX stems are considered to be an abstract way of representing a whole inflectional paradigm, inflectional affixes and inflected stems were not included in the database, and the token frequency associated with a stem is the sum of the token frequencies of the stem and all its inflected forms. However, MORPA also contains inflectional rules of which the token frequencies should be available. For obtaining frequency information on inflectional affixes and stems, we had to use the CELEX corpus, containing approximately 44 million words. Unfortunately, the morphological information in this corpus does not contain any production rules or information on the affixes.

Thus, after all production rules in CELEX had been counted straightforwardly, we were only able to assign frequency information to a part of the MORPA rules. Moreover, we knew that some of

[9] Although not in the grammar, this symbol is used to make it possible to describe the possibility of a word being of a certain category in terms of (5).

[10] For the reader's convenience, the probabilities denote the tree (in labelled bracketing) and production rules involved.

these frequencies were overestimated. Because we expected these facts to have a negative influence on the accuracy of the PCFG, we decided to put some effort in making the empirical estimate more reliable. We had to be very creative in finding other ways to provide the rules which are not in CELEX with frequency information (from CELEX), but we finally managed to provide almost all production rules employed by MORPA with frequency information. Also, we put some effort into "repairing" the overestimated frequencies. Consequently, the data have become more complete and more reliable, but as a result of these problems, the collection of frequency information became a time-consuming and error-sensitive job: a lot of work had to be done by hand. Therefore, it is practically almost undoable to go over it all over again.

With respect to the reliability of the frequency data, it turned out that the token frequencies are less reliable than the lexical frequencies. Most importantly, this was due to the fact that in CELEX, the token frequencies were "string" counts, i.e. they indicate how many times each separate string of letters occurs in the 44-million-word corpus. Because some of these "separate strings of letters" may be ambiguous in word class, morphemic segmentation or meaning, they are assigned different entries in the stems corpus. Ideally, the token frequencies in the corpus are disambiguated for the different entries, but at the time we collected our data they were not[11]. As a consequence, numerous stems were assigned overestimated token frequencies.

Consider, for example, the string *met*, which can be linked to two entries in the stems database: the entry of the preposition *met* 'with', and the entry of the noun *met* 'minced pork'. Since the individual frequencies of each of these entries have not been sorted out, the rules "$P \rightarrow$ met" and "$N \rightarrow$ met" have the same frequency, i.e. the frequency of the string *met*. Because the preposition is highly frequent and the noun hardly ever occurs, the latter rule has been assigned a frequency which is highly overestimated. Since in addition to that overestimation the rule "$w \rightarrow N$" is more frequent than the rule "$w \rightarrow P$", and to the frequency of the rule "$N \rightarrow$ met" is added the frequency of the two compounds in which it takes part, MORPA will consider the noun to be the most likely analysis. Had the frequencies been sorted out, this would not be the case: the high probability of the rule "$P \rightarrow$ met" would have overweighted all other probabilities.

The unreliability of token frequencies was beared out by some preliminary tests, in which we experimented using type and token frequencies on both dictionary and text test samples. When examining MORPA's output on a text test sample (for which token frequencies were used), we discovered that many of the erroneous selections were indeed attributable to the lack of disambiguation of token frequencies. Especially if the sample contained highly frequent string ambiguous simplex words, such as *met*, which do not take part in derivation or compounding, MORPA's performance got worse. It turned out that MORPA's performance was best, when type frequencies were used in a dictionary test sample.

MORPA first generates all possible parses and the associated probabilities, ordering them along a scale of plausibility afterwards. Thus, as yet, it is not a probabilistic parser in the sense that it prunes the low probability parses in an early stage [Fujisaki *et al.*, 1989; Jelinek *et al.*, 1990]. Adjusting the parser will speed it up considerably, but also pruning low-ranked analyses may lead to incompleteness.

In conclusion, let us return to the example word *beneveling*. After likelihood determination and ordering of the two remaining analyses in (8), the correct analysis be + nevel + ing is in topmost position:

(14) 1 be + nevel + ing N
2 be + neef + eling N

4 The performance of MORPA

In order to evaluate the performance of our system a test was run on a dictionary test sample of 3,077 words. The words contained in this sample were randomly taken from texts of the so-called "Bloemendal corpus" [Bringmann, 1990].

For a correct interpretation of the results, it is necessary to know that a word was considered to be correctly analysed, if it had been assigned the correct morphemic segmentation and word class. The analysis in (15) is the correct analysis of the word *beneveling*:

(15) [$_N$ [$_{prefix}$ be] [$_{stem}$ nevel] [$_{suffix}$ ing]]

Thus, in the final output of MORPA, morphological information which is irrelevant for pronunciation is eliminated: analyses which have the same segmentation, but are ambiguous in their hierarchical structure and/or categorial labelling of the morphemes, such as (5a) and (5b), become one as long as the morphemes have the same morphological classification, e.g. ((non)-native) prefix, suffix or stem, and the word is assigned the same word class.

As MORPA combines a conventional grammar with a probability-based scoring function, it is interesting to look at the effects of both the rule-based part and the probability-based ordering technique in their own right: the segmentation procedure and grammar determine the quality of the analyses and the number of analyses generated, and the probability-based

[11] By now, CELEX has disambiguated the token frequencies, but as the collection of reliable data was very time-consuming, we have not yet "repaired" our token frequencies.

scoring function enables MORPA to select the most likely analysis from a set of alternatives.

The results in (16) show how well the segmentation procedure and grammar succeeded in deriving the correct analysis for the test words:

(16)

words assigned	Number n = 3,077	%
a correct analysis	2,968	96
no correct analysis	32	1
no analysis at all	77	3

MORPA assigned no analysis at all to 3% of the test words. For 1% of the test words, one or more analyses were generated, but the set of alternatives did not contain a correct analysis. In these cases, the word either contains an unknown morpheme, or the grammar is too restrictive. 96% of the test words were assigned a correct analysis.

Given the problem of ambiguity, the number of analyses generated for one word is remarkably small: considering only the words which were correctly analysed, MORPA assigned a single, correct analysis to 46% of the test words. For 54%, the correct analysis was among alternatives:

(17)

words assigned a correct analysis, which is	Number n = 2,968	%
among alternatives	1,612	54
unique	1,356	46

Although we did not keep track of the number of segmentations assigned to the input words, it can be generally assumed that the number of alternative segmentations is very much reduced by the grammar. Also, through converting output that contains hierarchical structures and categorial labels (*cf.* (5)a and (5)b) to linear structures and morpheme classification (*cf.* (15)), a lot of unnecessary ambiguity is eliminated.

In order to evaluate the probability-based scoring function, which enables MORPA to order competing analyses along a scale of plausibility, it must be established how often MORPA succeeds in selecting the correct analysis from a set of alternatives. MORPA was able to select the correct analysis as most likely member of a set of alternatives for 92% of the test words. For a proper judgement of this performance, the percentage must be compared with the chances of selecting the most likely analysis from the set of alternatives. This chance is determined at 40%:

(18)

words assigned the best analysis from a set of alternatives, by	Number n = 1,612	%
the probability-based ordering technique	1,483	92
chance	645	40

It is not easy to tell which factors attributed to the fact that for 8% of the words the correct analysis was not selected as best analysis. The frequency data may be unreliable or the probability function may not be appropriate. Also, the correct analysis does not always have to be the most probable one.

Most importantly however, is the overall performance of MORPA's PCFG on the Bloemendal corpus: 92% of the test words had been assigned a correct analysis which was also the first analysis yielded.

(19)

words assigned	Number n = 3,077	%
a correct analysis in topmost position	2,835	92

For the 8% of the test words which were not assigned a correct analysis in first position, MORPA either generated a correct analysis which was not in first position, or no correct analysis or no analysis at all.

In order to establish the relevance for word level pronunciation, a test was run on a test file containing approximately 2000 isolated words. The test words were selected from different corpora to make sure the file contained both newspaper text, dictionary words

and words of frequency 1[12]. The words of the test file were analysed by MORPA and the topmost analyses were used by MORPHON to derive a pronunciation transcription. A transcription was considered correct if it had the proper phonemic transcription, which means that all appropriate non-optional phonological rules must have been applied, and that the words must have the correct syllable structure and stress pattern.

Fifteen percent of the words were assigned an erroneous phonemic transcription[13]. Twenty percent of the errors could be traced back to the phonological module, the remaining errors, 80%, are due to faulty morphological analyses. Of the errors made by MORPA, 88% led to an incorrect pronunciation representation. As expected, segmentation errors almost always led to an incorrect phonemic transcription. Category assignment errors also cause incorrect pronunciations, though less often. This bears out the importance of the category a word belongs to.

5 Conclusion

As the results show, this fully implemented system, running with a morpheme lexicon of 17,087 entries on a randomly selected 3,077 words test sample, is successful. This success may to a large extent be put down to the augmentation of the context-free grammar to a PCFG[14].

As mentioned above, the accuracy of a PCFG depends heavily on the accuracy of the empirical estimate of the probability function. We were lucky to have at our disposal a training set which was both large enough and representative, but due to the facts that, in some cases, MORPA and the training set yield different analyses, and token frequencies for string ambiguous words were not disambiguated, we expect our estimate to have become less reliable. In order to improve MORPA's performance on text test samples, we will have to "repair" the token frequencies.

It is often argued that a PCFG only provides poor estimates of probability, and that probabilistic grammars require more sensitivity to lexical context. After all, PCFGs only provide very general information on how likely a production rule is going to appear anywhere in a sample of the language, and production rules are not always context-free [Magerman and Marcus, 1991; Resnik, 1992]. However, most of the work done on context-free probabilistic grammars is done for syntax, and as I hope to have shown that a PCFG yields good results for morphology, it might be interesting to find out if, for one reason or another, PCFGs are more successful for morphology than for syntax.

[12] For reasons I will not go into here, the newspaper and dictionary words did not comprise highly frequent words [Nunn and van Heuven, 1993].

[13] See for a comparison with a data-oriented system for Dutch grapheme-to-phoneme transcription [van den Bosch and Daelemans, 1993]. Note that in this comparison syllabification and stress assignment have not been taken into account.

[14] Before this augmentation, the parser was enriched with some preliminary criteria imposing an order on the set of alternatives. Then, the performance came up to 85%.

Acknowledgements

I wish to thank my former colleagues of the Phonetics Laboratory at Leiden University who contributed to the work on MORPA. Furthermore, I am greatly indebted to Louis ten Bosch for his help with probability theory and Emiel Krahmer and Wessel Kraaij for solving all my LaTeX problems.

References

[Allen *et al.*, 1987] J. Allen, M.S. Hunnicutt, and D. Klatt. *From Text to Speech: the MITalk System.* Cambridge University Press, 1987.

[van Beurden, 1987] L. van Beurden. Playing level with Dutch morphology. In F. Beukema and P. Coopmans, editors, *Linguistics is the Netherlands 1987*, pages 21–30, 1987.

[van den Bosch and Daelemans, 1993] A. van den Bosch and W. Daelemans. Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics.*, 1993.

[Bringmann, 1990] E. Bringmann. Philip Bloemendal corpus. Internal Report 21, Analysis and Synthesis of speech, Utrecht 1990.

[Burnage, 1990] Gavin Burnage. *CELEX, a guide for users.* CELEX Centre for Lexical Information, Nijmegen, 1990.

[Chomsky and Halle, 1968] N. Chomsky and M. Halle. *The sound pattern of English.* Harper and Row , New York, 1968.

[Dirksen and Quené, 1993] A. Dirksen and H. Quené. Prosodic analysis: the next generation. In V. van Heuven and L. Pols, editors, *Analysis and Synthesis of Speech, Strategic Research Towards High-Quality Text-to-Speech Generation*, pages 131–145. Mouton de Gruyter, Berlin, 1993.

[Dowty, 1979] D. Dowty. *Word meaning and Montague Grammar.* Foris Dordrecht, 1979.

[Fujisaki *et al.*, 1989] T. Fujisaki, F. Jelinek, J. Cocke, E. Black, and T. Nishino. A probabilistic parsing method for sentence disambiguation. In *International Workshop on Parsing Technologies*, Pittsburgh P.A., 1989.

[Heemskerk and van Heuven, 1993] J. Heemskerk and V. van Heuven. MORPA: a

morpheme lexicon-based morphological parser. In V. van Heuven and L. Pols, editors, *Analysis and Synthesis of Speech, Strategic Research Towards High-Quality Text-to-Speech Generation.* Mouton de Gruyter, Berlin, 1993.

[Heemskerk, 1989] J. Heemskerk. Morphological parsing and lexical morphology. In H. Bennis and A. van Kemenade, editors, *Linguistics in the Netherlands 1989*, pages 61–70, 1989.

[Hoeksema, 1984] J. Hoeksema. *Categorial Morphology.* PhD thesis, Groningen, 1984.

[Jelinek *et al.*, 1990] F. Jelinek, J.D. Lafferty, and R.L. Mercer. Basic methods of probabilistic context free grammars. Research Report R.C. 16374(72684), IBM, 1990.

[van Leeuwen and te Lindert, 1993] H.C. van Leeuwen and E. te Lindert. Speech-Maker: a flexible framework for constructing text-to-speech sytems. In V. van Heuven and L. Pols, editors, *Analysis and Synthesis of Speech, Strategic Research Towards High-Quality Text-to-Speech Generation.* Mouton de Gruyter, Berlin, 1993.

[Liberman, 1991] M.J. Liberman. The trend toward statistical models in natural language processing. In E. Klein and F. Veltman, editors, *Natural Language and Speech.* Springer Verlag:Berlin, 1991.

[Magerman and Marcus, 1991] D. Magerman and M. Marcus. Pearl: A probabilistic chart parser. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, 1991.

[Moortgat, 1987] M. Moortgat. Compositionality and the syntax of words. In J. Groenendijk, D. de Jongh, and M. Stokhof, editors, *Foundations of Pragmatics and Lexical semantics*, 1987.

[Nunn and van Heuven, 1993] A. Nunn and V. van Heuven. MORPHON, lexicon-based text-to-phoneme conversion and phonological rules. In V. van Heuven and L. Pols, editors, *Analysis and Synthesis of Speech, Strategic Research Towards High-Quality Text-to-Speech Generation.* Mouton de Gruyter, Berlin, 1993.

[Resnik, 1992] P. Resnik. Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In *Proceedings International Conference on Computational Linguistics*, Nantes, 1992.

[Siegel, 1979] D. Siegel. *Topics in English Morphology.* Garland: New York, 1979.

[Spencer, 1991] A. Spencer. *Morphological Theory.* Basil Blackwell, 1991.

[van der Wouden, 1988] T. van der Wouden. Celex: Building a multifunctional polytheoretical lexical database. In *Proceedings Budalex*, 1988.