Memory-Based Morphological Analysis

Antal van den Bosch and **Walter Daelemans**
ILK / Computational Linguistics
Tilburg University
{antalb,walter}@kub.nl

Abstract

We present a general architecture for efficient and deterministic morphological analysis based on memory-based learning, and apply it to morphological analysis of Dutch. The system makes direct mappings from letters in context to rich categories that encode morphological boundaries, syntactic class labels, and spelling changes. Both precision and recall of labeled morphemes are over 84% on held-out dictionary test words and estimated to be over 93% in free text.

1 Introduction

Morphological analysis is an essential component in language engineering applications ranging from spelling error correction to machine translation. Performing a full morphological analysis of a wordform is usually regarded as a segmentation of the word into morphemes, combined with an analysis of the interaction of these morphemes that determine the syntactic class of the wordform as a whole. The complexity of wordform morphology varies widely among the world's languages, but is regarded quite high even in the relatively simple cases, such as English. Many wordforms in English and other western languages contain ambiguities in their morphological composition that can be quite intricate. General classes of linguistic knowledge that are usually assumed to play a role in this disambiguation process are knowledge of (i) the morphemes of a language, (ii) the morphotactics, i.e., constraints on how morphemes are allowed to attach, and (iii) spelling changes that can occur due to morpheme attachment.

State-of-the art systems for morphological analysis of wordforms are usually based on two-level finite-state transducers (FSTs, Koskenniemi (1983)). Even with the availability of sophisticated development tools, the cost and complexity of hand-crafting two-level rules is high, and the representation of concatenative compound morphology with continuation lexicons is difficult. As in parsing, there is a trade-off between coverage and spurious ambiguity in these systems: the more sophisticated the rules become, the more needless ambiguity they introduce.

In this paper we present a learning approach which models morphological analysis (including compounding) of complex wordforms as sequences of classification tasks. Our model, MBMA (Memory-Based Morphological Analysis), is a memory-based learning system (Stanfill and Waltz, 1986; Daelemans et al., 1997). Memory-based learning is a class of inductive, supervised machine learning algorithms that learn by storing examples of a task in memory. Computational effort is invested on a "call-by-need" basis for solving new examples (henceforth called instances) of the same task. When new instances are presented to a memory-based learner, it searches for the best-matching instances in memory, according to a task-dependent similarity metric. When it has found the best matches (the *nearest neighbors*), it transfers their solution (classification, label) to the new instance. Memory-based learning has been shown to be quite adequate for various natural-language processing tasks such as stress assignment (Daelemans et al., 1994), grapheme–phoneme conversion (Daelemans and Van den Bosch, 1996; Van den Bosch, 1997), and part-of-speech tagging (Daelemans et al., 1996b).

The paper is structured as follows. First, we give a brief overview of Dutch morphology in Section 2. We then turn to a description of MBMA in Section 3. In Section 4 we present the experimental outcomes of our study with MBMA. Section 5 summarizes our findings, reports briefly on a partial study of English showing that the approach is applicable to other languages, and lists our conclusions.

2 Dutch Morphology

The processes of Dutch morphology include inflection, derivation, and compounding. Inflection of verbs, adjectives, and nouns is mostly achieved by suffixation, but a circumfix also occurs in the Dutch past participle (e.g. **ge+werk+t** as the past participle of verb **werken**, to work). Irregular inflectional morphology is due to relics of ablaut (vowel change) and to suppletion (mixing of different roots in inflectional paradigms). Processes of derivation in Dutch morphology occur by means of prefixation and suffixation. Derivation can change the syntactic class of wordforms. Compounding in Dutch is concatenative (as in German and Scandinavian languages): words can be strung together almost unlimitedly, with only a few morphotactic constraints, e.g., **rechtsinformatica-toepassingen** (applications of computer science in Law). In general, a complex wordform inherits its syntactic properties from its right-most part (the head). Several spelling changes occur: apart from the closed set of spelling changes due to irregular morphology, a number of spelling changes is predictably due to morphological context. The spelling of long vowels varies between double and single (e.g. **ik loop**, I run, versus **wij lop+en**, we run); the spelling of root-final consonants can be doubled (e.g. **ik stop**, I stop, versus **wij stopp+en**, we stop); there is variation between **s** and **z** and **f** and **v** (e.g. **huis**, house, versus **huizen**, houses). Finally, between the parts of a compound, a linking morpheme may appear (e.g. **staat+s+loterij**, state lottery). For a detailed discussion of morphological phenomena in Dutch, see De Haas and Trommelen (1993). Previous approaches to Dutch morphological analysis have been based on finite-state transducers (e.g., XEROX'es morphological analyzer), or on parsing with context-free word grammars interleaved with exploration of possible spelling changes (e.g. Heemskerk and van Heuven (1993); or see Heemskerk (1993) for a probabilistic variant).

3 Applying memory-based learning to morphological analysis

Most linguistic problems can be seen as context-sensitive mappings from one representation to another (e.g., from text to speech; from a sequence of spelling words to a parse tree; from a parse tree to logical form, from source language to target language, etc.) (Daelemans, 1995). This is also the case for morphological analysis. Memory-based learning algorithms can learn mappings (classifications) if a sufficient number of instances of these mappings is presented to them.

We drew our instances from the CELEX lexical data base (Baayen et al., 1993). CELEX contains a large lexical data base of Dutch wordforms, and features a full morphological analysis for 247,415 of them. We took each wordform and its associated analysis, and created task instances using a windowing method (Sejnowski and Rosenberg, 1987). Windowing transforms each wordform into as many instances as it has letters. Each example focuses on one letter, and includes a fixed number of left and right neighbor letters, chosen here to be five. Consequently, each instance spans eleven letters, which is also the average word length in the CELEX data base. Moreover, we estimated from exploratory data analysis that this context would contain enough information to allow for adequate disambiguation.

To illustrate the construction of instances, Table 1 displays the 15 instances derived from the Dutch example word **abnormaliteiten** (abnormalities) and their associated classes. The class of the first instance is "A+Da", which says that (i) the morpheme starting in **a** is an adjective ("A")[1], and (ii) an **a** was deleted at the end ("+Da"). The coding thus tells that the first morpheme is the adjective **abnormaal**. The second morpheme, **iteit**, has class "N_A∗". This complex tag indicates that when **iteit** attaches right to an adjective (encoded by "A∗"), the new combination becomes a noun ("N_"). Finally, the third morpheme is **en**, which is a plural inflection (labeled "m" in CELEX). This way we generated an instance base of 2,727,462

[1] CELEX features ten syntactic tags: noun (N), adjective (A), quantifier/numeral (Q), verb (V), article (D), pronoun (O), adverb (B), preposition (P), conjunction (C), interjection (J), and abbreviation (X).

instances. Within these instances, 2422 different class labels occur. The most frequently occurring class label is "0", occurring in 72.5% of all instances. The three most frequent non-null labels are "N" (6.9%), "V" (3.6%), and "m" (1.6%). Most class labels combine a syntactic or inflectional tag with a spelling change, and generally have a low frequency.

When a wordform is listed in CELEX as having more than one possible morphological labeling (e.g., a morpheme may be N or V, the inflection **-en** may be plural for nouns or infinitive for verbs), these labels are joined into ambiguous classes ("N/V") and the first generated example is labeled with this ambiguous class. Ambiguity in syntactic and inflectional tags occurs in 3.6% of all morphemes in our CELEX data.

The memory-based learning algorithm used within MBMA is IB1-IG (Daelemans and Van den Bosch, 1992; Daelemans et al., 1997), an extension of IB1 (Aha et al., 1991). IB1-IG constructs a data base of instances in memory during learning. New instances are classified by IB1-IG by matching them to all instances in the instance base, and calculating with each match the *distance* between the new instance X and the memory instance Y, $\Delta(X,Y) = \sum_{i=1}^{n} W(f_i)\delta(x_i,y_i)$, where $W(fi)$ is the weight of the ith feature, and $\delta(x_i,y_i)$ is the distance between the values of the ith feature in instances X and Y. When the values of the instance features are symbolic, as with our linguistic tasks, the simple *overlap* distance function δ is used: $\delta(x_i,y_i) = 0\ if\ x_i = y_i,\ else\ 1$. The (most frequently occurring) classification of the memory instance Y with the smallest $\Delta(X,Y)$ is then taken as the classification of X.

The weighting function $W(f_i)$ computes for each feature, over the full instance base, its *information gain*, a function from information theory; cf. Quinlan (1986). In short, the information gain of a feature expresses its relative importance compared to the other features in performing the mapping from input to classification. When information gain is used in the similarity function, instances that match on important features are regarded as more alike than instances that match on unimportant features.

In our experiments, we are primarily interested in the *generalization accuracy* of trained models, i.e., the ability of these models to use their accumulated knowledge to classify new instances that were not in the training material. A method that gives a good estimate of the generalization performance of an algorithm on a given instance base, is 10-fold cross-validation (Weiss and Kulikowski, 1991). This method generates on the basis of an instance base 10 subsequent partitionings into a training set (90%) and a test set (10%), resulting in 10 experiments.

4 Experiments: MBMA of Dutch wordforms

As described, we performed 10-fold cross validation experiments in an experimental matrix in which MBMA is applied to the full instance base, using a context width of five left and right context letters. We structure the presentation of the experimental outcomes as follows. First, we give the generalization accuracies on test instances and test words obtained in the experiments, including measurements of generalization accuracy when class labels are interpreted at lower levels of granularity. While the latter measures give a rough idea of system accuracy, more insight is provided by two additional analyses. First, precision and recall rates of morphemes are given. We then provide prediction accuracies of syntactic word classes. Finally, we provide estimations on free-text accuracies.

4.1 Generalization accuracies

The percentages of correctly classified test instances are displayed in the top line of Table 2, showing an error in test instances of about 4.1% (which is markedly better than the baseline error of 27.5% when guessing the most frequent class "0"), which translates in an error at the word level of about 35%. The output of MBMA can also be viewed at lower levels of granularity. We have analyzed MBMA's output at the three following lower granularity levels:

1. Only decide, per letter, whether a segmentation occurs at that letter, and if so, whether it marks the start of a derivational stem or an inflection. This can be derived straightforwardly from the full-task class labeling.
2. Only decide, per letter, whether a segmentation occurs at that letter. Again, this can

instance number	left context					focus letter	right context					TASK
1	_	_	_	_	_	**a**	**b**	**n**	**o**	**r**	**m**	A+Da
2	_	_	_	_	**a**	**b**	**n**	**o**	**r**	**m**	**a**	0
3	_	_	_	**a**	**b**	**n**	**o**	**r**	**m**	**a**	**l**	0
4	_	_	**a**	**b**	**n**	**o**	**r**	**m**	**a**	**l**	**i**	0
5	_	**a**	**b**	**n**	**o**	**r**	**m**	**a**	**l**	**i**	**t**	0
6	**a**	**b**	**n**	**o**	**r**	**m**	**a**	**l**	**i**	**t**	**e**	0
7	**b**	**n**	**o**	**r**	**m**	**a**	**l**	**i**	**t**	**e**	**i**	0
8	**n**	**o**	**r**	**m**	**a**	**l**	**i**	**t**	**e**	**i**	**t**	0
9	**o**	**r**	**m**	**a**	**l**	**i**	**t**	**e**	**i**	**t**	**e**	N_A*
10	**r**	**m**	**a**	**l**	**i**	**t**	**e**	**i**	**t**	**e**	**n**	0
11	**m**	**a**	**l**	**i**	**t**	**e**	**i**	**t**	**e**	**n**	_	0
12	**a**	**l**	**i**	**t**	**e**	**i**	**t**	**e**	**n**	_	_	0
13	**l**	**i**	**t**	**e**	**i**	**t**	**e**	**n**	_	_	_	0
14	**i**	**t**	**e**	**i**	**t**	**e**	**n**	_	_	_	_	m
15	**t**	**e**	**i**	**t**	**e**	**n**	_	_	_	_	_	0

Table 1: Instances with morphological analysis classifications derived from **abnormaliteiten**, analyzed as **[abnormaal]**$_A$**[iteit]**$_{N\_A*}$**[en]**$_m$.

be derived straightforwardly. This task implements segmentation of a complex word form into morphemes.

3. Only check whether the desired spelling change is predicted correctly. Because of the irregularity of many spelling changes this is a hard task.

The results from these analyses are displayed in Table 2 under the top line. First, Table 2 shows that performance on the lower-granularity tasks that exclude detailed syntactic labeling and spelling-change prediction is about 1.1% on test instances, and roughly 10% on test words. Second, making the distinction between inflections and other morphemes is almost as easy as just determining whether there is a boundary at all. Third, the relatively low score on correctly predicted spelling changes, 80.95%, indicates that it is particularly hard to generalize from stored instances of spelling changes to new ones. This is in accordance with the common linguistic view on spelling-change exceptions. When, for instance, a past-tense form of a verb involves a real exception (e.g., the past tense of Dutch **brengen**, to bring, is **bracht**), it is often the case that this exception is confined to generalize to only a few other examples of the same verb (**brachten**, **gebracht**) and not to any other word that is not derived from the same stem, while the memory-based learning approach is not aware of such constraints. A post-processing step that checks whether the proposed morphemes are also listed in a morpheme lexicon would correct many of these errors, but has not been included here.

4.2 Precision and recall of morphemes

Precision is the percentage of morphemes predicted by MBMA that is actually a morpheme in the target analysis; recall is the percentage of morphemes in the target analysis that are also predicted by MBMA. Precision and recall of morphemes can again be computed at different levels of granularity. Table 3 displays these computed values. The results show that both precision and recall of fully-labeled morphemes within test words are relatively low. It comes as no surprise that the level of 84% recalled fully labeled morphemes, including spelling information, is not much higher than the level of 80% correctly recalled spelling changes (see Table 2). When word-class information, type of inflection, and spelling changes are discarded, precision and recall of basic segment types becomes quite accurate: over 94%.

class labeling granularity	labeling example	instances %	instances ±	words %	words ±
full morphological analysis	**[abnormaal]**$_A$**[iteit]**$_{N\_A*}$**[en]**$_m$	95.88	0.04	64.63	0.24
derivation/inflection	**[abnormal]**$_{deriv}$**[iteit]**$_{deriv}$**[en]**$_{infl}$	98.83	0.02	89.62	0.17
segmentation	**[abnormal][iteit][en]**	98.97	0.02	90.69	0.02
spelling changes	+Da	80.95	0.40	—	—

Table 2: Generalization accuracies in terms of the percentage of correctly classified test instances and words, with standard deviations (±) of MBMA applied to full Dutch morphological analysis and three lower-granularity tasks derived from MBMA's full output. The example word **abnormaliteiten** is shown according to the different labeling granularities, and only its single spelling change at the bottom line).

task variation	precision (%)	recall (%)
full morphological analysis	84.33	83.76
derivation/inflection	94.72	94.07
segmentation	94.83	94.18

Table 3: Precision and recall of morphemes, derived from the classification output of MBMA applied to the full task and two lower-granularity variations of Dutch morphological analysis, using a context width of five left and right letters.

4.3 Predicting the syntactic class of wordforms

Since MBMA predicts the syntactic label of morphemes, and since complex Dutch wordforms generally inherit their syntactic properties from their right-most morpheme, MBMA's syntactic labeling can be used to predict the syntactic class of the full wordform. When accurate, this functionality can be an asset in handling unknown words in part-of-speech tagging systems. The results, displayed in Table 4, show that about 91.2% of all test words are assigned the exact tag they also have in CELEX (including ambiguous tags such as "N/V" – 1.3% wordforms in the CELEX dataset have an ambiguous syntactic tag). When MBMA's output is also considered correct if it predicts at least one out of the possible tags listed in CELEX, the accuracy on test words is 91.6%. These accuracies compare favorably with a related (yet strictly incomparable) approach that predicts the word class from the (ambiguous) part-of-speech tags of the two surrounding words, the first letter, and the final three letters of Dutch words, viz. 71.6% on unknown words in texts (Daelemans et al., 1996a).

syntactic class prediction	correct test words (%)	words ±
exact	91.24	0.21
exact or among alternatives	91.60	0.21

Table 4: Average prediction accuracies (with standard deviations) of MBMA on syntactic classes of test words. The top line displays exact matches with CELEX tags; the bottom line also includes predictions that are among CELEX alternatives.

4.4 Free text estimation

Although some of the above-mentioned accuracy results, especially the precision and recall of fully-labeled morphemes, seem not very high, they should be seen in the context of the test they are derived from: they stem from held-out portions of dictionary words. In texts sampled from real-life usage, words are typically smaller and morphologically less complex, and a relatively small set of words re-occurs very often. It is therefore relevant for our study to have an estimate of the performance of MBMA on real texts. We generate such an estimate following these considerations: New, unseen text is bound to contain a lot of words that are in the 245,000 CELEX data base, but also some number of unknown words. The morphological analyses of known words are simply retrieved by the memory-based learner from memory. Due to some ambiguity in the class labeling in the data base itself, retrieval accuracy will be somewhat below 100%. The morphological analyses of unknown words are assumed to be as accurate as was tested in the above-mentioned experiments: they can be said to be of the type of dictionary words in the 10% held-out test sets of 10-fold cross validation experiments. CELEX bases its wordform frequency information on word counts made on the 42,380,000-words Dutch INL corpus. 5.06% of these wordforms are wordform tokens that occur only once. We assume that this can be extrapolated to the estimate that in real texts, 5% of the words do not occur in the 245,000 words of the CELEX data base. Therefore, a sensible estimate of the accuracies of memory-based learners on real text is a weighted sum of accuracies comprised of 95% of the *reproduction accuracy* (i.e, the error on the training set itself), and 5% of the generalization accuracy as reported earlier.

Table 5 summarizes the estimated generalization accuracy results computed on the results of MBMA. First, the percentages of correct instances and words are estimated to be above 98% for the full task; in terms of words, it is estimated that 84% of all words are fully correctly analyzed. When lower-granularity classification tasks are discerned, accuracies on words are estimated to exceed 96% (on instances, less than 1% errors are estimated). Moreover, precision and recall of morphemes on the full task are estimated to be above 93%. A considerable surplus is obtained by memory retrieval in the estimated percentage of correct spelling changes: 93%. Finally, the prediction of the syntactic tags of wordforms would be about 97% according to this estimate.

We briefly note that Heemskerk (1993) reports a correct word score of 92% on free text test material yielded by the probabilistic morphological analyzer MORPA. MORPA segments wordforms, decides whether a morpheme is a stem, an affix or an inflection, detects spelling changes, and assigns a syntactic tag to the wordform. We have not made a conversion of our output to Heemskerk's (1993). Moreover, a proper comparison would demand the same test data, but we believe that the 92% corresponds roughly to our MBMA estimates of 97.2% correct syntactic tags, 93.1% correct spelling changes, and 96.7% correctly segmented words.

Estimate	value
correct instances, full task	98.4%
correct words, full task	84.2%
correct instances, derivation/inflection	99.6%
correct words, derivation/inflection	96.7%
correct instances, segmentation	99.6%
correct words, segmentation	96.7%
precision of fully-labeled morphemes	93.6%
recall of fully-labeled morphemes	93.2%
precision of deriv./infl. morphemes	98.5%
recall of deriv./infl. morphemes	98.0%
precision of segments	98.5%
recall of segments	97.9%
correct spelling changes	93.1%
exactly correct syntactic wordform tags	97.2%

Table 5: Estimations of accuracies on real text, derived from the generalization accuracies of MBMA on full Dutch morphological analysis.

5 Conclusions

We have demonstrated the applicability of memory-based learning to morphological analysis, by reformulating the problem as a classification task in which letter sequences are classified as marking different types of morpheme boundaries. The generalization performance of memory-based learning algorithms to the task is encouraging, given that the tests are done on held-out (dictionary) words. Estimates of free-text performance give indications of high accuracies: 84.6% correct fully-analyzed words (64.6% on unseen words), and 96.7% correctly segmented and coarsely-labeled words (about 90% for unseen words). Precision and recall of fully-labeled morphemes is estimated in real texts to be over 93% (about 84% for unseen words). Finally, the prediction of (possibly ambiguous) syntactic classes of unknown wordforms in the test material was shown to be 91.2% correct; the corresponding free-text estimate is 97.2% correctly-tagged wordforms.

In comparison with the traditional approach, which is not immune to costly hand-crafting and spurious ambiguity, the memory-based learning approach applied to a reformulation of the problem as a classification task of the segmentation type, has a number of advantages:

- it presupposes no more linguistic knowledge than explicitly present in the corpus used for training, i.e., it avoids a knowledge-acquisition bottleneck;
- it is language-independent, as it functions on any morphologically analyzed corpus in any language;
- learning is automatic and fast;
- processing is deterministic, non-recurrent (i.e., it does not retry analysis generation) and fast, and is only linearly related to the length of the wordform being processed.

The language-independence of the approach can be illustrated by means of the following partial results on MBMA of English. We performed experiments on 75,745 English wordforms from CELEX and predicted the lower-granularity tasks of predicting morpheme boundaries (Van den Bosch et al., 1996). Experiments yielded 88.0% correctly segmented test words when deciding only on the location of morpheme boundaries, and 85.6% correctly segmented test words discerning between derivational and inflectional morphemes. Both results are roughly comparable to the 90% reported here (but note the difference in training set size).

A possible limitation of the approach may be the fact that it cannot return more than one possible segmentation for a wordform. E.g. the compound word **kwartslagen** can be interpreted as either **kwart+slagen** (quarter turns) or **kwarts+lagen** (quartz layers). The memory-based approach would select one segmentation. However, true segmentation ambiguity of this type is very rare in Dutch. Labeling ambiguity occurs more often (3.6% of all morphemes), and the current approach simply produces ambiguous tags. However, it is possible for our approach to return distributions of possible classes, if desired, as well as it is possible to "unpack" ambiguous labeling into lists of possible morphological analyses of a wordform. If, for example, MBMA's output for the word **bakken** (bake, an infinitive or plural verb form, or bins, a plural noun) would be $\textbf{[bak]}_{V/N}\textbf{[en]}_{tm/i/m}$, then this output could be expanded unambiguously into the noun analysis $\textbf{[bak]}_{N}\textbf{[en]}_{m}$ (plural) and the two verb readings $\textbf{[bak]}_{V}\textbf{[en]}_{i}$ (infinitive) and $\textbf{[bak]}_{V}\textbf{[en]}_{tm}$ (present tense plural).

Points of future research are comparisons with other morphological analyzers and lemmatizers; applications of MBMA to other languages (particularly those with radically different morphologies); and qualitative analyses of MBMA's output in relation with linguistic predictions of errors and markedness of exceptions.

Acknowledgements

This research was done in the context of the "Induction of Linguistic Knowledge" (ILK) research programme, supported partially by the Netherlands Organization for Scientific Research (NWO). The authors wish to thank Ton Weijters and the members of the Tilburg ILK group for stimulating discussions. A demonstration version of the morphological analysis system for Dutch is available via ILK's homepage `http://ilk.kub.nl`.

References

D. W. Aha, D. Kibler, and M. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

R. H. Baayen, R. Piepenbrock, and H. van Rijn. 1993. *The CELEX lexical data base on CD-ROM.* Linguistic Data Consortium, Philadelphia, PA.

W. Daelemans and A. Van den Bosch. 1992. Generalisation performance of backpropagation learning on a syllabification task. In M. F. J. Drossaers and A. Nijholt, editors, *Proc. of TWLT3: Connectionism and Natural Language Processing*, pages 27–37, Enschede. Twente University.

W. Daelemans and A. Van den Bosch. 1996. Language-independent data-oriented grapheme-to-phoneme conversion. In J. P. H. Van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, editors, *Progress in Speech Processing*, pages 77–89. Springer-Verlag, Berlin.

W. Daelemans, S. Gillis, and G. Durieux. 1994. The acquisition of stress: a data-oriented approach. *Computational Linguistics*, 20(3):421–451.

W. Daelemans, J. Zavrel, and P. Berck. 1996a. Part-of-speech tagging for Dutch with MBT, a memory-based tagger generator. In K. van der Meer, editor, *Informatiewetenschap 1996, Wetenschappelijke bijdrage aan* *de Vierde Interdisciplinaire Onderzoeksconferentie Informatiewetenchap*, pages 33–40, The Netherlands. TU Delft.

W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. 1996b. MBT: A memory-based part of speech tagger generator. In E. Ejerhed and I. Dagan, editors, *Proc. of Fourth Workshop on Very Large Corpora*, pages 14–27. ACL SIGDAT.

W. Daelemans, A. Van den Bosch, and A. Weijters. 1997. IGTree: using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review*, 11:407–423.

W. Daelemans. 1995. Memory-based lexical acquisition and processing. In P. Steffens, editor, *Machine Translation and the Lexicon*, Lecture Notes in Artificial Intelligence, pages 85–98. Springer-Verlag, Berlin.

W. De Haas and M. Trommelen. 1993. *Morfologisch handboek van het Nederlands: Een overzicht van de woordvorming.* SDU, 's Gravenhage, The Netherlands.

J. Heemskerk and V. van Heuven. 1993. 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 speech generation*, pages 67–85. Mouton de Gruyter, Berlin.

J. Heemskerk. 1993. A probabilistic context-free grammar for disambiguation in morphological parsing. In *Proceedings of the 6th Conference of the EACL*, pages 183–192.

K. Koskenniemi. 1983. *Two-level morphology: a general computational model for word-form recognition and production.* Ph.D. thesis, University of Helsinki.

J.R. Quinlan. 1986. Induction of Decision Trees. *Machine Learning*, 1:81–206.

T. J. Sejnowski and C. S. Rosenberg. 1987. Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168.

C. Stanfill and D. Waltz. 1986. Toward memory-based reasoning. *Communications of the* ACM, 29(12):1213–1228, December.

A. Van den Bosch, W. Daelemans, and A. Weijters. 1996. Morphological analysis as classification: an inductive-learning approach. In K. Oflazer and H. Somers, editors, *Proceedings of the Second International Conference on New Methods in Natural Language Processing, NeMLaP-2, Ankara, Turkey*, pages 79–89.

A. Van den Bosch. 1997. *Learning to pronounce written words: A study in inductive language learning.* Ph.D. thesis, Universiteit Maastricht.

S. Weiss and C. Kulikowski. 1991. *Computer systems that learn.* San Mateo, CA: Morgan Kaufmann.