Evaluation and Adaptation of the Celex Dutch Morphological Database

Tom Laureys*, Guy De Pauw†, Hugo Van hamme*,
Walter Daelemans†, Dirk Van Compernolle*

*K.U.Leuven/ESAT/PSI
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
{tom.laureys,hugo.vanhamme,dirk.vancompernolle}@esat.kuleuven.ac.be

†University of Antwerp/CNTS
Universiteitsplein 1, 2610 Antwerpen, Belgium
{guy.depauw,walter.daelemans}@ua.ac.be

Abstract

This paper describes some important modifications to the Celex morphological database in the context of the FLaVoR project. FLaVoR aims to develop a novel modular framework for speech recognition, enabling the integration of complex linguistic knowledge sources, such as a morphological model. Morphology is a fairly unexploited linguistic information source speech recognizers could benefit from. This is especially true for languages which allow for a rich set of morphological operations, such as our target language Dutch. In this paper we focus on the exploitation of the Celex Dutch morphological database as the information source underlying two different morphological analyzers being developed within the project. Although the Celex database provides a valuable source of morphological information for Dutch, many modifications were necessary before it could be practically applied. We identify major problems, discuss the implemented solutions and finally experimentally evaluate the effect of our modifications to the database.

1. Introduction

This paper describes some important modifications to the Celex Dutch morphological database that transform it into a readily applicable information source for a modular speech recognition engine. These modifications were deemed paramount for exploiting this type of morphological information during the recognition process, but more generally also help in providing a more consistent and more widely applicable morphological database.

Although it is generally acknowledged that more accurate linguistic knowledge sources (phonology, morphology, syntax) are crucial for improving speech recognition accuracy, truly powerful language models have seldom been incorporated into speech recognizers (Rosenfeld, 2000). The main reason is that the standard recognition architecture requires all knowledge sources to be extremely simple as it combines them all into one single search space. In the FLaVoR project (Flexible Large Vocabulary Recognition) we try to overcome this restriction by means of a novel, more flexible speech recognition architecture which splits the search engine into two separate layers: a layer for acoustic-phonemic decoding and one for word decoding. This way, more complex linguistic information can be applied in the word decoding step (Demuynck et al., 2003).

One valuable linguistic information source that has recently been applied in speech recognition is morphology. A recognizer can benefit from morphological information in two major ways. First, the use of morphemes allows for both a reduction of the lexicon size and of the number of out-of-vocabulary (OOV) words. Second, morphological features can be integrated into the recognizer's language model to improve accuracy. It is clear that modeling morphology is especially important for speech recognition of morphologically rich languages. Successful experiments have recently been reported for Finnish (Siivola et al., 2003) and Hungarian (Szarvas and Furui, 2003). For Dutch, encouraging results have already been obtained by explicitly modeling compounds (Laureys et al., 2002).

Within the FLaVoR project two very different approaches to this type of morphological analysis are being developed in parallel, however both requiring a substantial amount of Dutch morphological training data. Currently, Celex is the only extensive and publicly available morphological database for Dutch (Baayen et al., 1995). Unfortunately, this database is not readily applicable as an information source in a practical system due to both a considerable amount of annotation errors and a number of practical considerations. The research described in this paper attempts to identify these problems and resolve them in a systematic way. The effect of the adaptations was evaluated in a small-scale experiment.

The paper is structured as follows. First, we sketch our view on the application of morphology in speech recognition. Next, we describe the Celex Dutch morphological database. Then, the necessary adaptations are discussed and experimentally evaluated. We conclude with some suggestions for future work.

2. Speech Recognition and Morphology

2.1. FLaVoR Architecture

In the standard speech recognition framework all knowledge sources are applied as early as possible in the search process. The main advantage of this approach is the efficiency of the search: early inclusion of higher level information from the lexicon and the language model is beneficial for reliably pruning away the most unlikely hypotheses. Yet, at the same time this architecture forces all knowledge sources to be extremely simple in structure and to preferably operate from left to right. As a result, these (linguistic) models can only be crude approximations (e.g. N-

grams, lexicalization).

In the FLaVoR project a novel framework is proposed which splits the search into two layers:

- The first layer gets acoustic features as input and outputs a phoneme network.
- The second layer performs the actual word decoding starting from the phoneme network. Two important knowledge sources operate in this layer: a morpho-phonological model which converts the phoneme network into corresponding sequences of morphemes (with word boundary hypotheses), and a morpho-syntactic language model.

In figure 1 the standard recognition architecture and FLaVoR's architecture are compared.

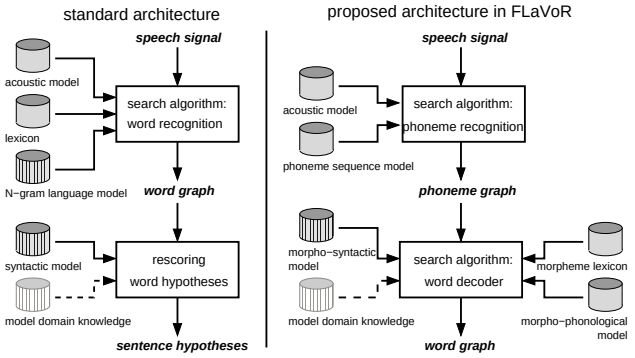

Figure 1: Standard vs. FLaVoR architecture

2.2. Morphology in FLaVoR

The morpho-syntactic language model will provide a probability measure for each hypothesized word based on morphological and syntactic information of the word and its context. In the current phase of the project, research focuses on the morphological component, whichs aims to provide a hierarchical morphological analysis (covering flection, derivation and compounding) for each input word and an associated probability measure for the analysis.

To emphasize the modular approach of the FlaVoR project, two different morphological analyzers are being developed in parallel, based on different techniques:

- an analyzer based on machine learning techniques (memory-based learning) which defines morphological analysis as a classification problem;
- an analyzer based on finite state transducers, extended with probabilistic post-processing.

Although they are quite dissimilar in approach, both systems require a rich morphological training source for the induction of instances, rules and probabilities. Likewise, a test set needs to be extracted from that information source for an objective comparison between the two systems. Unfortunately, no such readily applicable morphological information source is available, making it necessary to adapt an existing morphological database to our purpose.

3. The Celex Dutch Morphological Database

Celex is currently the only extensive, publicly available morphological knowledge source for Dutch. It contains 381.292 word forms corresponding to 124.136 headwords.

The headword lexicon provides a detailed hierarchical analysis for each entry. This analysis identifies all morphemes the entry is composed of and allows to draw a complete morphological tree diagram for the word. In addition, each morpheme is assigned a tag: roots receive their part-of-speech tag, affixes get a tag indicating their combinatorial status. For example, the following is the hierarchical analysis for the entry *onmisbaar* [E indispensable]:
`((on)[A|.A],((mis)[V],(baar)[A|V.])[A])[A]`
For the affix tags, the letter combination following the vertical bar refers to the part-of-speech tags of the morpheme(s) the affix is combined with, while the single letter in front of the bar refers to the part-of-speech tag of the result. The dot indicates the position of the affix. In the example above, the verbal stem *mis* is first combined with the suffix *baar*. Then, the resulting adjective *misbaar* is prefixed with *on*.

In the word forms lexicon each entry is linked to its corresponding entry in the headword lexicon. In addition, inflectional information is provided by means of features (e.g. number, person, case, etc.). The adaptations we will discuss in the next segment all pertain to the hierarchical analyses in the headword lexicon.

4. Adaptations to Celex

First, we discuss the major obstacles for exploiting the Celex morphological database in the morphological analyzers developed in the FLaVoR project. After we provide solutions to these problems, we detail the adaptation procedure itself and provide quantitative results.

4.1. Problematic Issues

Despite the amount of detailed morphological analyses for Dutch, Celex in its original shape is inadequate as an information source within the FLaVoR framework. Two main reasons can be discerned. First, we found that the Celex morphological database contains a considerable number of annotation errors, inconsistencies and missing or dubious analyses. This can be explained by the size of the database which forced the developers to use automatic annotation systems, combined with partial manual checking. Second, Celex has approached morphology from the viewpoint of theoretical linguistics. In some cases, however, the resulting analyses are no longer compatible with the practical requirements of the recognition architecture as a whole. We will discuss both types of problems.

4.1.1. Annotation Inaccuracies

Many annotation errors were retrieved by examining low frequency events (morphemes, tags, morpheme combinations, etc.) and by comparing the output of our own analyzers-under-development with the Celex analysis. We list the major types of analyses we deemed inaccurate. They represent 6.2% (7646 items) of the original database.

The headword lexicon contains a number of flected forms (plurals, diminutives, participles), which we would have expected to be confined to the word forms lexicon. In

addition, those forms are treated inaccurately, as they are just mapped to the form without flection affixes. For example, the diminutive *sterretje* [E small star] is analyzed as its headword *ster* [E star], i.e. ignoring the diminutive suffix.

Many compounds and derivatives are left unanalyzed. This type of flaw can only be traced systematically by comparing the output of an automatic analyzer with the Celex analysis. Consider the following examples: *aardewerk* [E earthenware], *filmscript* [E film script], *socialist* [E socialist], *absurdisme* [E absurdism], etc.

Further, a number of non-systematic inconsistencies or errors were resolved. For example, although *blauwzijden* [E blue silk], *roodzijden* [E red silk] and *witzijden* [E white silk] have a morphologically analogous structure, they are inconsistently analyzed.

```
((blauw)[A],(zijde)[N],(en)[A|AN.])[A]
((rood)[A],((zijde)[N],(en)[A|N.])[A])[A]
(witzijden)[A]
```

Examples of plain errors are the analyses of *kwikkuur* [E mercury cure] and *papier* [E paper]:

```
((kwik)[N],(uur)[N|N.])[N]
```
and
```
((paap)[N],(ier)[N|N.])[N]
```
instead of
```
((kwik)[N],(kuur)[N])[N]
```
and
```
(papier)[N]
```

4.1.2. Practical Adaptations

Practical considerations forced us to adapt another 7.6% (9476 items) of the entries which are correct from a theoretical point of view, but do not fit into the proposed recognition framework.

The main motivation is that Celex uses many truncation operations. Yet, in the recognition framework described in section 2., truncation is incompatible with the morpho-phonological model as it entails hypothesizing morphemes which are not acoustically realized. For example, *epiek* [E epic] is analyzed as:

```
(((epos)[A],(isch)[N|A.])[N],(iek)[N|N.])[N]
```

This analysis first involves truncation of the ending *-os*[1] and the adjectival suffix *-isch*. We avoided the use of truncation by the introduction of bound morphemes which are roots (i.e. not affixes). The use of these bound morphemes is standard in Dutch morphology and allows us to model the regular morphological processes in loan words (De Haas and Trommelen, 1993). For *epiek* the resulting analysis is:

```
((ep)[G],(iek)[N|G.])[N]
```

The tag G refers to bound root morphemes. Clearly, the latter representation allows for an easier mapping between phonemes and morphemes and also enables a more transparant production and analysis of derivatives like the correpsonding adjective *episch* [E epic].

Other practical adaptations involved the quasi etymological analysis Celex provides for acronyms and abbreviations. For example, *bieb*, the abbreviation for *bibliotheek* [E library], was originally analyzed as its full form. Finally, non-standard orthographical alternations used by Celex were adapted. For example, Celex analyzes *platboomd* [E flat-bottomed] as follows:

```
((plat)[A],(bodem)[N],(d)[A|AN.])[A]
```

Yet, the orthographic conversion from *bodem* into *boom* is by no means standard. We chose to lexicalize such cases.

4.2. Adaptation Procedure and Effects

The complete adaptation procedure was implemented by means of scripts. Each of the scripts carries out a specific adaptation. This has three advantages:

- it is easy to keep track of the set of adaptations;
- similar adaptations can be performed by means of regular expressions;
- at each point, erroneous adaptations can easily be undone.

A careful manual check of the list of adaptations was necessary to avoid errors introduced mainly by 'overgreedy' regular expressions.

The adaptations had an effect on the number of different morphemes in the database: the analyses in the original version of Celex contained 269.789 morphemes, while the adapted version seems to provide more detailed analyses, consisting of a total of up to 279.009 morphemes. But whereas the initial corpus held 34.293 unique morphemes, the adapted version reduces this number to 32.727 unique morphemes[2]. This means that the productive power of the morpheme lexicon has risen since more detailed analyses are being generated by fewer morphemes. So despite the reduction in the number of morphemes we are confident that the generative capacity of the lexicon has been extended, especially by the introduction of 1616 bound root morphemes.

5. Experimental Evaluation

When modifying a large database, there is a risk that one introduces errors or loses consistency by making changes to certain word forms while neglecting analogous other ones. Therefore, we tried to measure the effect of the adaptations mentioned above by means of a small-scale experiment. The type of experiment is not strictly watertight, but still provides a good indication of the consistency of analyses throughout the database.

5.1. Experimental Setup

The experimental setup we used was largely inspired by (Van den Bosch and Daelemans, 1999), in which a memory-based learner is trained and tested on the Celex Dutch morphological database. It redefines morphological analysis as a classification task involving local decisions on the level of the grapheme. Instances consist of each letter in the lemma, its surrounding context and its associated morphological classification. In (Van den Bosch and Daelemans, 1999), this classification not only includes morpheme boundaries, but also part-of-speech tag, allomorphy and truncation information. The system is able to model complex morphological processes with a high degree of accuracy.

[1] The ending *-os* has a Greek origin. It is not a productive Dutch suffix.

[2] Tags are taken into account.

Context L					F	Context R					B
-	-	-	-	-	**a**	r	b	e	i	d	-
-	-	-	-	a	**r**	b	e	i	d	s	-
-	-	-	a	r	**b**	e	i	d	s	f	-
-	-	a	r	b	**e**	i	d	s	f	i	-
-	a	r	b	e	**i**	d	s	f	i	l	-
a	r	b	e	i	**d**	s	f	i	l	o	+
r	b	e	i	d	**s**	f	i	l	o	s	+
b	e	i	d	s	**f**	i	l	o	s	o	-
e	i	d	s	f	**i**	l	o	s	o	o	-
i	d	s	f	i	**l**	o	s	o	o	f	-
d	s	f	i	l	**o**	s	o	o	f	i	-
s	f	i	l	o	**s**	o	o	f	i	e	-
f	i	l	o	s	**o**	o	f	i	e	-	-
i	l	o	s	o	**o**	f	i	e	-	-	-
l	o	s	o	o	**f**	i	e	-	-	-	+
o	s	o	o	f	**i**	e	-	-	-	-	-
s	o	o	f	i	**e**	-	-	-	-	-	+

Table 1: Instances for the morphological analysis of *arbeidsfilosofie* [E work philosophy]

To properly evaluate the adaptations on Celex, we implemented a simplified version of this system that can predict morpheme boundaries in the concatenation of the morphemes of a morphological analysis (i.e. no orthographic alternation rules were implemented). This reduces the classification task significantly by not requiring the analyzer to hypothesize part-of-speech tag information and the like. An example of training instances derived from the Dutch word *arbeidsfilosofie* [E work philosophy] can be seen in table 1. F refers to the focus letter, B to the presence of a morpheme boundary. Our setup also enables us to evaluate the classification task as a pattern-matching task. We hypothesize that a more consistently annotated corpus should facilitate this type of pattern-matching task.

5.2. Discussion of Results

	Instances	Morphemes	Words
Celex (original)	97.7%	90.7%	82.4%
Celex (adapted)	98.6%	92.8%	87.4%

Table 2: Experimental results

The results in table 2 show that the system trained and tested on the adapted version of Celex is indeed able to achieve a much higher accuracy on our classification task. Almost 99% of all instances were correctly classified in the adapted version. Also the F-score[3] on the morpheme level is significantly higher. On the word level we notice an increase up to 87% of correctly analyzed words. These results indicate that the higher degree of consistency has a positive effect on the classification accuracy of our basic morphological analyzer, despite the fact that the analyses have become more detailed and therefore arguably harder to predict. Despite the relatively simple setup of the experiments described above, it seems clear that the new version of Celex benefits from the higher degree of consistency. The higher annotation accuracy for the database should consequently reflect itself in better morphological analyzers derived from it.

[3] A weighted average of the standard precision and recall metrics.

6. Conclusions and Future Work

Future work will include a fully realized version of the memory-based analyzer, as well as an independently developed system for morphological analysis using finite state techniques. A detailed comparison will be made of the two systems sharing the same data set, which will enable us to identify their respective strengths and weaknesses. It is expected that this line of research, alongside experiments with active learning and typicality measures, can identify more annotation errors and inconsistencies in Celex. But the adjustments described in this paper already go a long way in turning Celex into a more adequate information source for data-driven morphological analysis in Dutch.

7. Acknowledgements

The research described in this paper was funded by IWT in the GBOU programme, project FLaVoR: Flexible Large Vocabulary Recognition: Incorporating Linguistic Knowledge Sources Through a Modular Recogniser Architecture.(Project number 020192). http://www.esat.kuleuven.ac.be/spch/projects/FLaVoR.

8. References

Baayen, R.H., R. Piepenbrock, and L. Gulikers, 1995. *The Celex Lexical Database (Release2) [CD-ROM]*. Philadelphia, U.S.A.: Linguistic Data Consortium, University of Pennsylvania.

De Haas, W. and M. Trommelen, 1993. *Morfologisch Handboek van het Nederlands: een Overzicht van de Woordvorming*. 's Gravenhage, The Netherlands: SDU.

Demuynck, K., T. Laureys, D. Van Compernolle, and H. Van hamme, 2003. Flavor: a flexible architecture for LVCSR. In *Proc. European Conference on Speech Communication and Technology*. Geneva, Switzerland.

Laureys, T., V. Vandeghinste, and J. Duchateau, 2002. A hybrid approach to compounds in LVCSR. In *Proc. International Conference on Spoken Language Processing*, volume I. Denver, U.S.A.

Rosenfeld, R., 2000. Two decades of statistical language modeling: Where do we go from here? *Proc. of the IEEE*, 88(8):1270–1278.

Siivola, V., T. Hirsimaki, M. Creutz, and M. Kurimo, 2003. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In *Proc. European Conference on Speech Communication and Technology*. Geneva, Switzerland.

Szarvas, M. and S. Furui, 2003. Finite-state transducer based modeling of morphosyntax with applications to Hungarian LVCSR. In *Proc. International Conference on Acoustics, Speech and Signal Processing*. Hong Kong, China.

Van den Bosch, A. and W. Daelemans, 1999. Memory-based morphological analysis. In *Proc. 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*. New Brunswick, U.S.A.