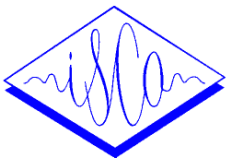

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

A LEXICON FOR A TEXT-TO-SPEECH SYSTEM

Léon Gulikers and Rijk Willemse
Nijmegen University,
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

ABSTRACT

This paper describes the development of a full-form lexicon, combined with an algorithm for quasi-morphological decomposition aiming at improved grapheme-to-phoneme conversion, word stress assignment, syllabification and word class assignment in a Text-to-Speech system. We will explain the way in which the optimal size of the lexicon was determined. Also, we describe a deterministic algorithm for decomposing words not found in the lexicon in terms of a sequence of lexicon entries and prefixes, suffixes and infixes. The performance of the lexicon+decomposition system is evaluated with a newspaper corpus comprising approximately 100,000 words. It appears that the system handles more than 95% of the regular words in the test corpus correctly. The system will have to be extended with a module that handles proper names.

INTRODUCTION

Advanced Text-to-Speech (TTS) systems for unlimited input text require substantial linguistic processing. Intonation and Prosody require detailed knowledge of the surface structure of the sentences that make up the text. Ideally, semantic interpretation, as well as topic-comment or given-new information across sentences should also be available. Moreover, in most western languages correct grapheme-to-phoneme conversion and word stress assignment requires knowledge about the morphological structure of the words.

The need for a (large) lexicon for a syntactic parser is obvious. However, Dutch is one of the languages that pose big problems for the developer of a lexicon, because the conventional spelling system of the language promotes the use of compounds in which the parts are not separated by blanks, as is usually the case in English. Consequently, if 'words' are defined as 'sequences of alphabetic characters not containing blanks', the lexicon of the Dutch language is infinite for all practical applications involving unrestricted text. Moreover, morphological decomposition of words not found in a finite lexicon is far from trivial. Many words appear to be decomposable into a large number of morpheme structures, most of which do not make sense, unfortunately for so many different reasons that it has not yet been possible to keep over-generation within reasonable bounds. Last but not least, many ambiguous morphological analyses appear to lead to the same syntactic class of the compound. For a purely class-based syntactic analysis this is, of course, no problem. However, there is a number of words for which grapheme-to-phoneme conversion or word stress assignment depends on the morphological structure. In these cases one would need semantic interpretation to disambiguate the structures. For a TTS system for unrestricted text semantic interpretation is obviously far beyond the present capabilities.

In the past, two approaches have been used for grapheme–to–phoneme conversion for Dutch. The first one [3] is completely based on letter-to-sound rules. Detailed analysis of the rules in this system shows that they use a lot of quasi-morphological knowledge; for instance, many rules insert 'morpheme' boundaries in consonant clusters or after a closed set of prefixes. Although the rule based system performs reasonably well, it appears to fail on a large set of not-so-frequent words, most of which appear to be compounds that defy the morphology embedded in the rules. The only way to improve the performance of the rules seems to be the addition of information about morpheme boundaries [4].

The other approach starts with a complete morphological analysis of the words in the input text [2]. The analysis is based on a comprehensive list of morphemes occurring in Dutch, plus a large grammar with its attendant parser. The backtrack parser generates a number of possible analyses, that are ordered by means of a number of simple heuristics. The morpheme lexicon contains phonemic representations; phonemic forms of complex words are computed from these abstract phonemic base forms plus a small number of rules. The actual performance of this system is somewhat disappointing. The morphological parser still fails to find correct parses for a nonnegligible number of words in unrestricted text, if only because it cannot handle foreign words. Equally disturbing is the fact that the backtrack parser may take more than half an hour CPU time on a VAX 3100 workstation, only to produce a number of wrong analyses of longish words.

These observations led us to the decision to build a TTS system that is based on a large (but finite) full-form lexicon, combined with deterministic morphological analysis of words not appearing in the lexicon. Since it is our aim to build a practical TTS system, it is not only important that the analyses of the words in an input sentence are correct, but also that the system can approach real-time performance on a PC-like computer. Also, the storage capacity needed for the lexicon must remain within reasonable bounds. Both the processing time and the memory size restrictions will contribute to limiting the actual coverage of the lexicon on unlimited text. Thus, it is clear that we have to search for an optimal compromise solution.

THE SIZE OF THE LEXICON

The optimal size of the lexicon in our system must be based on a compromise between effective coverage of arbitrary Dutch texts on the one hand and restrictions on processing time and storage capacity on the other.

Several lexica, with sizes ranging between 12,000 and 24,000 words were tested with respect to their coverage. Of course, frequency of occurrence of the words in a number of corpora was among the most important criteria for inclusion of a word in a lexicon [1]. The first lexicon comprises all words that have a frequency > 200 in the CELEX [1] lexicon for Dutch. The frequency counts in this lexicon are based on a 42 million word corpus, provided by the Institute for Dutch Lexicography in Leiden (INL). This corpus is mainly based on fiction texts. The 'CELEX' lexicon of high frequency words comprised approximately 12,000 entries.

The second lexicon was also derived from a list that was based on frequency of occurrence. In the early eighties frequency of occurrence of word forms was counted in the Dutch regional newspaper 'de Haarlemse Courant' on behalf of the University of Utrecht [5]. The resulting lexicon is known as the ULEX lexicon. The frequency counts in ULEX are based on one and a half million words taken from regular articles. Advertisements were not included in the counting. In order to be comparable with the CELEX lexicon, the ULEX lexicon in our experiment was formed by taking the 12,000 most frequent words.

A third lexicon was created by merging the CELEX and ULEX lexicons. The resulting lexicon, which we will indicate as C+U, contains about 16,000 words.

Since we knew that it would be necessary to enhance the lexicon look-up with some form of morphological decomposition, we decided to create a fourth lexicon, consisting of 'monomorphematic' words. This lexicon was again created with the help of the CELEX lexical data base, viz. by taking all words from that data base which do not contain a marker for a morpheme boundary. The quotes around the term monomorphematic reflect the theoretical problem in deciding whether a word in a given language must be considered as simple or complex. CELEX treats loanwords that may be considered as complex in the language of origin but that function as simple words in Dutch as monomorphematic. The monomorphematic lexicon comprised 20,800 words.

Finally, a fifth lexicon was created by merging the C+U lexicon with the monomorphematic lexicon. The most important characteristics of the five lexicons are summarized in Table 1.

[1] CELEX is the CEntre for LEXical information in Nijmegen

10.21437/ICSLP.1992-28

Table 1: possible lexicons

lexicon	
CELEX	12,000 most frequent words
ULEX	12,000 most frequent words
C+U	Combination of CELEX and ULEX 16,000 words
Monomorphemic	20,800 words
C+U+M	Combination of CELEX, ULEX and the monomorphemic words 24,500 words

To find the best lexicon, two test-corpora were used:

- Gelderland corpus
- General corpus

The 'Gelderland' corpus consist of 123,000 words taken from the regional newspaper 'de Gelderlander'. The issues used for our research appeared in the late eighties. Also, the newspaper from which ULEX was derived has no relation whatsoever to 'the Gelderlander'. The 'General' corpus was composed of the following texts:

- the "Eindhoven" corpus, that consists of a large variety of text samples, taken from newspaper, magazines, popular science books, etc,.
- short stories of the writer F.B. Hotz
- legal texts published by the European Community
- texts from the newspaper 'de Gelderlander'; there was no overlap between these texts and the texts of the 'Gelderlander' corpus mentioned earlier.

The 'General' corpus consists of 1,5 million words.

Figure 1 shows the result of an experiment in which the two test corpora were matched against the five lexicons. 'Matching' was not limited to straightforward look-up; a simple mechanism for decomposing test words that were not found in the lexicons was part of the look-up procedure. Essentially, the decomposition was limited to attempts to build 'new' words by concatenating words that did occur in the lexicon. By doing so, a fair proportion of nominal compounds can be found. At this stage, the algorithm for morphological decomposition did not include prefix and suffix stripping. (A more detailed description of the

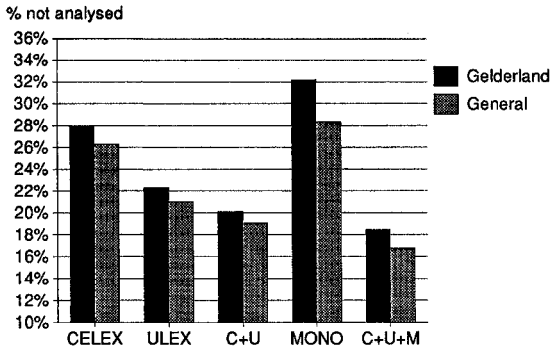

Figure 1: Analyses with possible lexicons

eventual algorithm can be found in the Section 'Morphological Decomposition'.)

Figure 1 shows that although the CELEX and ULEX lexicons both consist of the 12,000 most frequent words in the text they were derived from, there is a substantial difference in the proportion of the words in the test corpus that they cover. The fact that the frequency counts in the CELEX data base are based on a corpus that consists mainly of fiction, probably explains why the coverage of the CELEX lexicon is so much lower than that of the ULEX lexicon, especially when the 'Gelderland' corpus is tested. It is well known that fictional texts differ substantially from other types of texts in that they tend to have a higher type-token ratio; i.e., fiction texts contain more different words, and few words have a high frequency.

Figure 1 also shows that adding the list of monomorphematic words to the C+U lexicon makes almost no difference, except that it increases the size of the lexicon a lot. It is not quite clear why the performance increase is negligible. There are at least two possibilities: once again, things would improve when a more powerful algorithm for morphological decomposition were used, or the extra morphemes have a very low frequency of occurrence, so that they are unlikely to improve the coverage significantly.

After what has been said before, the low performance of the lexicon comprising only monomorphematic forms is no longer surprising. It can only be a good basic lexicon when it is combined with a powerful morphological analysis, and that is by and large what we would like to avoid in order to prevent excessive computation and excessive over-generation. Our tentative explanations of the causes for the different performance of the five lexica are supported by the results of a test that was run with a later, improved version of the morphological decomposition. Although the number of analyses using the lexicon with monomorphematic words were higher than shown in Fig. 1, they still were not as good as the results of analyses with the C+U lexicon. Whereas the C+U lexicon reached a coverage of a little more than 80% on the 'de Gelderlander' corpus, only 74% of the words in that corpus could be analysed with the lexicon of monomorphematic words.

It is well known that there is no such thing as 'the' coverage of a lexicon; each domain of the world addressed in a specific text will require a specially tuned lexicon in order to maximize the coverage. Yet, the results of the coverage study suggest that in most cases the C+U lexicon will show the best coverage.

Before embarking upon experiments aiming at optimization of the morphological analysis the C+U lexicon was processed in order to remove a number of words that we think should not be covered by a general purpose basic lexicon. For instance, all proper names were removed from the lexicon. We think that names should (and can) be handled by a separate module [7].

Together with the names, foreign words were also removed from the lexicon, unless they had a very high frequency of occurrence. A small number of foreign words was retained because they have become so established that they take part in active morphological processes, like compounding and derivation. Also, all abbreviations were removed from the C+U lexicon, for the same reason: they should be handled by a separate module, if only because the expansion of abbreviations in Dutch depends to a large extent on the domain of the world addressed by the text. Numbers were also removed, with the exception of words like, 'one', ..., 'ten', 'hundred', 'thousand', etc. Finally, a number of word forms were removed that can be generated by simple morphological rules, like plural and diminutive forms of nouns. In those cases were the plural or diminutive form of a word was present, but not the base form, the derived form was replaced by the base form.

THE RECORDS IN THE LEXICON

One of the biggest advantages of lexicon-based text-to-speech conversion is the low rate of errors that are made during the morpho-phonological part of TTS conversion, because the phoneme representation, syllable structure and information about stress are correctly given in the lexicon.

Although relatively few cases occur, Dutch does possess a number of heterophone homographs. One such example is the word "onderwijs"; with stress on "on", the word is a noun, meaning education, while with stress on "wijs" it is the first person singular form of a verb meaning to educate. Whenever two homographs have different pronunciation depending on the word class of the word, we had to link the pronunciation to the word class. In those cases where heterophone homographs have the same syntactic class, the best we can do is to select the pronunciation connected to the meaning that occurs most frequently.

Improvement of the morpho-phonological part of TTS conversion was one of the goals that we wanted to achieve. In the TTS system that we are building a lexicon is also an indispensible component of the syntactic analysis. Since our syntactic analysis is completely de-lexicalised (i.e., it only operates on a very limited set of syntactic categories) a high performance Word Class Assignment is mandatory. In [6] the probabilistic algorithm for

word class assignment used in our system is described. Its major input is information on the probability that a word has one of its possible classes. This information is derived from tagged corpora, and then included in the lexicon.

MORPHOLOGICAL DECOMPOSITION

It will be clear that it is impossible to find all the words of a text, just by straightforward lexicon look-up. To find the remaining words, an algorithm for morphological decomposition will try to analyse them by using the following three procedures.

- prefix stripping
- suffix stripping
- splitting of compounds into the separate parts

Prefix stripping means that the algorithm tries to detect and strip prefixes from the beginning of a word. After a prefix has been stripped, the remaining part of the word form will be processed. Processing does not just mean lexicon look-up; the stripped word form will be handled as if it were found in the input text. Thus, if the stripped word is not found in the lexicon, it will be subjected to prefix stripping in its turn. The same recursion strategy is, of course, used for the two remaining analysis procedures. Though we call it prefix stripping, the list of 'prefixes' in our system contains not only regular prefixes like "on"(un) and "ont" (dis), but also a number of prepositions. The reason for this is that a lot of words tend to begin with either a prefix or one or more prepositions. The verb "achternalopen" (to follow) begins with the prepositions "achter" and "na" (the combination of these two can best be translated as "behind") and it ends with the stem "lopen" (to walk). By stripping the two prefixes first, the word can be analysed as a combination of the two prefixes and a stem.

Suffix stripping is almost the same process as prefix stripping; the only difference is that this part of the morphological decomposition deals with word endings formed by flexion and derivation.

Derivation of the word "moeilijk" (difficult) can lead to the word "moeilijkheid" (difficulty). If the word "moeilijkheid" occurs in a text, it will be analysed as the stem "moeilijk" and the suffix "heid".

Analysis of compounds in Dutch is an extremely complicated process, due to the fact that the spelling system promotes glueing together words without separating blanks. As already said in the introduction, many words appear to be decomposable into a large number of morpheme structures. One illustrative example is the word "speellokaal" (games room). The correct analyses would be "speel" (to play) and "lokaal" (room), but analyses like "spe" (future), "el" (yard), "lok" (entice) and "aal" (eel) are also possible. Using a non-deterministic algorithm, that would make a complete morphological analyses of this word, would lead a lot of possible analyses. To avoid these problems, our deterministic algorithm tries to decompose words by searching for the longest word that matches the right hand part of a complex word. To that end, a succesively growing number of characters is taken from the ending of the word, and looked up in the lexicon. Each time such a sequence of characters appears to be a word, the remaining left hand part of the input is analysed, in order to see if it too is a (simple or complex) word. If that is the case, the analysis is saved on a stack, but the process of growing the number of characters to be taken from the right hand side of the input word continues. Each time a new match is found, the previous one can be discarded. In the example word "speellokaal" the algorithm will successively find the words "al" (already), "aal", "kaal" (bald) and "lokaal". It will retain the analysis "speel" + "lokaal", which happens to be the correct one.

If the left hand part would appear to be a compound, like in the word "slaapkamerraam" (bedroom window), first "raam" would be found as the longest right hand part of the complete word. In searching for the left hand side, the algorithm would find "kamer" as the longest possible right hand part of the compound "slaapkamer". Again, this happens to be the correct analysis. Of course, this simple heuristic does not always provide the correct result.

ANALYSIS RESULTS

To test the performance of combination of the lexicon and the deterministic algorithm for morphological decomposition, in terms of the coverage of the lexicon, we used a newspaper corpus of 97,460 words (tokens), taken from the regional newspaper "de Gelderlander". Table 2 shows the results [2] of this test.

Table 2: Performance test

Total number of words	97,460
Words found in the lexicon	78,274 (80.31%)
Words found by decomposition	10,914 (11.19%)
Words missed	8,272 (8.48%)

Table 2 shows that 91.5% of the words of this test corpus either occur in the lexicon, or can be analysed as a sequence of simpler words from the lexicon. 8.5% of the words in the corpus cannot be handled by our algorithm. Although this proportion seems to be rather high, a better look at the words that were missed shows that it is impossible to find or to analyse 6,055 of these words, since they are:

- Abbreviations
- Names
- Foreign words
- Numbers

All these types of words were deliberately excluded from the lexicon. If a good text preprocessor is available for our TTS system, abbreviations and numbers will no longer cause problems, since a text preprocessor can convert them into words that can be found in the lexicon. As long as names are not the first word of a sentence, they can be recognized because of the capitalization; in this way it is at least possible to give them a proper word class. Rules for computing the correct pronunciation still await development.

It will not be easy to find a solution for foreign words, even if that would be very desirable. In many social circles, the linguistic culture tends to fully accept of the use of foreign words. In some newspapers, these foreign words can be distinguished, because they appear italicized. Unfortunately, this is the exception rather than the rule. But even if a word can be recognized as a loanword, it cannot be pronounced before its language of origin has been established. To accomplish that on a word by word basis is almost impossible. To understand that, it suffices to look at the spelling of words from French, English and German, the three languages from which most of the foreign words are taken. Especially for the less frequent words, it is extremely difficult to find sequences of two or more letters that have a much higher frequency in one language than in the other two. Today, most foreign words in Dutch texts tend to be English, a language known for the complex relation between spelling and pronunciation.

From the date given above it appears that there is a hard core (2.27%) of regular Dutch words that cannot be handled by our lexicon-plus-morphology system. We will need a rule-based letter-to-sound system to convert these words. Also, rules are needed to propose realistic syntactic categories for these words.

From the 10,914 words that were found by decomposing them into simpler entities, 188 (0.19% of the entire corpus) were assigned an incorrect analysis. Additional rules will be able to decrease the number of wrong analyses, but for the examples given below, there are probably no rules that can repair them. The only conceivable way to handle these words correctly is by including them in the lexicon, that therewith will tend to keep

[2] During this test we only looked at the phoneme representation that was found in the lexicon or built by combining the separate parts of a compound, that was analysed by the decomposition algorithm. Word classes can be assigned to the separate parts, but the final result of the word class assignment depends on the performance of the word class assignment module as it is described by Willemse and Gulikers [6].

increasing in size.

met+af+oor	(with + off + ear instead of metaphor)
tel+oor+gang	(count + ear + hall instead of loss)
best+reed	(best + drove instead of fought against)

One might think that the word 'bestreed' can be handled correctly if the process of prefix stripping (that would remove 'be') has precedence over the process that tries to decompose the word as a sequence of lexicon entries. Although true in principle, this is deceptive: changing the precedence relation between the processes appears to give rise to a higher number of errors than we have now.

If a word is found by decomposition, we need rules to compute the syllable that is to receive major word stress. E.g. if we were looking for the word "onderwijsdeskundige" (education expert), that will be analysed as "onderwijs" (education) and "deskundige"(expert), the syllables that can receive stress are,"onderwijs" and "deskundige". We will need rules that are able to detect that "onderwijs" is a specifier of "deskundige", and will therefore have primary stress, whereas "deskundige" can receive secondary stress. Fortunately, these rules are not difficult to formulate. Syntactic class of the parts of the compounds as well as the class of the compound itself play a decisive role. Class information is either present in the lexicon, or can be obtained from Word Class Assignment.

Rules are also necessary for the computation of the correct phonemic form of the compounds. In order to simplify these rules, the phonemic representations of all words in the lexicon are fairly abstract. For instance, word final devoicing of plosives and fricatives is not already accounted for in the phoneme forms in the lexicon. These simple processes are taken care of by a small set of rules. Another set of rules is needed for cross word assimilation processes. In our system, these rules also take care of cross-morpheme assimilation.

CONCLUSIONS

The lexicon as we have built it, proves to be a good alternative for TTS systems that solely work with rules for those parts that handle grapheme-to-phoneme conversion, stress assignment and syllabification, or systems that use non-deterministic techniques for morphological decomposition. However, even our approach can not handle all the words, especially not if we intend to analyse unrestricted texts. We therefore need additional rules to support the tasks of a TTS system that were just mentioned. One of the biggest improvements of our system, is the fact that this part of the TTS system needs very little time due to efficient routines for lexicon access and deterministic morphological decomposition.

Although we did not mention the issue very often in this paper, word class information of the words included in the lexicon is of great importance during further processing in the TTS system.

References

[1] J. Allen. M.S. Hunnicutt, D. Klatt. From text to speech: the MITalk system, Cambridge University Press, Cambridge, 1987.

[2] J.S.M. Heemskerk. MORPA, a morphological parser for use in a Dutch text-to-speech system. *SPIN/ASSP-report, in preparation. Speech Technology Foundation*, Utrecht, 1991.

[3] J. Kerkhoff, J. Wester and L. Boves. A compiler for implementing the linguistic phase of a text-to-speech conversion system, *Linguistics in the Netherlands*, H. Bennis & W.U.S. van Lessen Kloeke (eds), 111-117, 1984.

[4] J. Kerkhoff. Ph.D. Dissertation, to appear, 1992

[5] M.P.R. van den Broecke. Ter Sprake: spraak als betekenisvol geluid in 36 thematische hoofdstukken. *Floris publications*, 400-407, Dordrecht Holland/Providence RI USA, 1988

[6] Willemse, R. & Gulikers, L. Word class assignment in a text-to-speech system. Elsewhere in these proceedings.

[7] M.F. Spiegel, M.J. Macchi, K.D. Gollhardt, Synthesis of names by a demisyllable-based speech synthesizer (Spokesman), *Proceedings Eurospeech* 117-120, 1989