Lexical Embedding in Spoken Dutch

Odette Scharenborg¹ and Stefanie Okolowski²

¹ Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
² University of Trier, Germany

O.Scharenborg@let.ru.nl

Abstract

A stretch of speech is often consistent with multiple words, e.g., the sequence ‘ham/er’ is consistent with ‘ham’ but also with the first syllable of ‘hamster’, resulting in temporary ambiguity. However, to what degree does this lexical embedding occur? Analyses on two corpora of spoken Dutch showed that 11.9%-19.5% of polysyllabic word tokens have word-initial embedding, while 4.1%-7.5% of monosyllabic word tokens can appear word-initially embedded. This is much lower than suggested by an analysis of a large dictionary of Dutch. Speech processing thus appears to be simpler than one might expect on the basis of statistics on a dictionary.

Index Terms: lexical embedding, spoken-word recognition

1. Introduction

During spoken-word recognition, words that are consistent with the acoustic signal are activated and compete, after which the intended word sequence usually is recognised. However, spoken input is often consistent with multiple words, e.g., the phoneme sequence ‘/hæm/’ is consistent with the monosyllabic word ‘ham’, but also with the first syllable of ‘hamster’. Laboratory studies have shown that listeners use durational cues in the acoustic signal to resolve the temporary ambiguity due to lexical embedding (e.g., [1-3]). For instance, it is shown that the lexical interpretation of an embedded sequence is related to its duration; in carefully produced speech, a longer sequence tends to be interpreted as a monosyllabic word more often than a shorter one (e.g., [1,3]).

But to what degree does lexical embedding occur? This has been investigated using a dictionary for English [4] and has been shown to be substantial: a majority of polysyllabic words have shorter words embedded within them; moreover, these embedded words are most likely to appear at the onsets of their matrix words. However, dictionary analyses can only provide a rough indication of the size of the problem in real speech as words do not have equal chance of being encountered in real speech. Therefore, [5] carried out analyses on a real-speech corpus for English. They showed that 71.1% of all polysyllabic word tokens contained at least one embedded word, and again the majority of these embedded words appeared word-initially.

English, however, is a morphologically complex language such as Dutch. An analysis of a dictionary of Dutch showed that, also for Dutch, the majority of polysyllabic words have shorter words embedded, and that these embedded words are most likely to appear word-initially [6]. However, lexical embedding has not been investigated on real-speech. Therefore, in this paper, we investigate the degree to which lexical embedding occurs in spoken Dutch. To that end, we investigate the degree of lexical embedding in two corpora of spoken Dutch, containing multiple speaking styles. Since no study on lexical embedding in spoken Dutch has been carried out, the analyses will provide hitherto unknown information about the nature and structure of lexical embedding in everyday speech of Dutch.

2. Method

2.1. Analysing lexical embedding

We investigate lexical embedding by comparing the phonemic transcriptions of words. We first investigate lexical embedding in a large dictionary of Dutch, thus extending the work presented in [6], but using a much larger data set. We subsequently investigate lexical embedding using two corpora of spoken Dutch. The results of the analyses are split out in terms of length of the polysyllabic matrix words. Furthermore, we investigate the part-of-speech (POS) tags of the embedded and matrix words. Listeners might have expectations on the basis of POS information. For instance, when hearing a partial sentence like ‘ik zie een ham…’ (I see a ham…), it is most likely that ‘ham’ is the start of a noun, not a verb.

By looking at the phoneme level, the problem of lexical embedding might be overestimated. For example, it is possible that listeners use allophonic variation due to syllable structure during speech recognition. This issue is partly dealt with by only investigating lexical embeddings where the syllable boundaries of the matrix word match the embedded word. This thus excludes embeddings such as ‘boek’ (book) in ‘boeken’ (books), as the syllable boundary in the latter is before the /k/.

The analyses will thus present a ‘worst-case’ scenario.

2.2. Materials

The TST-lexicon is a Dutch dictionary consisting of 361,162 words, their phonemic transcriptions, word stress and syllable information, POS tags, and frequency of occurrence in the Spoken Dutch Corpus (CGN) [7]. It was compiled by merging lexical resources such as CELEX, Van Dale Dictionary of Dutch, and CGN. Single-phoneme words, words that are only used as parts of contracted multi-word expressions (e.g., ‘in- en uitvoer’, import and export), incomplete words, foreign words, and contracted words (e.g., ‘da’s’, that’s), and words with a frequency of zero in the CGN (71.2%) were excluded from the analyses. This yielded a set of 92,196 words. Furthermore, homophones (7.64%) were collapsed (ignoring the POS tags); this resulted in 85,150 different words.

For the analyses of lexical embedding in spoken Dutch, we used two corpora: the Northern Dutch part of the ‘core corpus’ of the CGN and the ‘informal’ and ‘story-retelling’ parts of the IFA corpus [8]. The core corpus of CGN consists of 675,417...
words, divided over 14 different speech styles/components (see Table 1). For each word, a manually verified phonemic transcription is available as well as a manually verified word-initial embedding; length 1 for Emb and Wi denotes the % word types that were embedded.

Table 4. TST statistics: the monosyllabic word types that can appear embedded and % polysyllabic word types with embedding that have at least one embedded word.

<table>
<thead>
<tr>
<th>Type of speech</th>
<th>#words</th>
<th>% mono</th>
<th>% mono emb</th>
<th>%P+emb</th>
<th>%P 2syll</th>
<th>%P 3syll</th>
<th>%P 4syll</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spont. conv. (face-to-face)</td>
<td>106,182</td>
<td>7.6</td>
<td>75.3</td>
<td>5.5</td>
<td>25.1</td>
<td>75.9</td>
<td>18.9</td>
</tr>
<tr>
<td>Interv. w/Dutch teachers</td>
<td>25,687</td>
<td>11.3</td>
<td>71.0</td>
<td>5.1</td>
<td>18.2</td>
<td>71.5</td>
<td>21.5</td>
</tr>
<tr>
<td>Spont. telephone dialog.</td>
<td>201,141</td>
<td>5.1</td>
<td>75.8</td>
<td>7.3</td>
<td>37.8</td>
<td>70.9</td>
<td>23.7</td>
</tr>
<tr>
<td>Sim. business negotiations</td>
<td>25,485</td>
<td>8.4</td>
<td>72.1</td>
<td>4.8</td>
<td>17.2</td>
<td>73.4</td>
<td>18.3</td>
</tr>
<tr>
<td>Interviews/discussions (B)</td>
<td>75,106</td>
<td>10.1</td>
<td>65.5</td>
<td>7.3</td>
<td>19.5</td>
<td>68.4</td>
<td>22.0</td>
</tr>
<tr>
<td>Political interv/disc. (nB)</td>
<td>25,117</td>
<td>12.6</td>
<td>59.5</td>
<td>7.2</td>
<td>13.6</td>
<td>57.1</td>
<td>31.3</td>
</tr>
<tr>
<td>Lessons in classroom</td>
<td>25,961</td>
<td>13.3</td>
<td>67.0</td>
<td>7.9</td>
<td>23.5</td>
<td>67.0</td>
<td>25.1</td>
</tr>
<tr>
<td>Live commentaries (B)</td>
<td>24,986</td>
<td>12.1</td>
<td>64.3</td>
<td>7.0</td>
<td>16.2</td>
<td>66.1</td>
<td>23.9</td>
</tr>
<tr>
<td>News reports (B)</td>
<td>25,065</td>
<td>15.2</td>
<td>64.5</td>
<td>7.1</td>
<td>18.8</td>
<td>63.1</td>
<td>27.4</td>
</tr>
<tr>
<td>News (B)</td>
<td>25,296</td>
<td>21.6</td>
<td>49.8</td>
<td>12.4</td>
<td>16.6</td>
<td>47.4</td>
<td>34.2</td>
</tr>
<tr>
<td>Comment/columns (B)</td>
<td>25,071</td>
<td>19.0</td>
<td>61.1</td>
<td>8.0</td>
<td>17.1</td>
<td>67.0</td>
<td>21.8</td>
</tr>
<tr>
<td>Speeches/sermons</td>
<td>5,184</td>
<td>21.1</td>
<td>60.3</td>
<td>6.7</td>
<td>12.7</td>
<td>67.7</td>
<td>23.4</td>
</tr>
<tr>
<td>Lectures/seminars</td>
<td>14,913</td>
<td>17.0</td>
<td>63.0</td>
<td>5.8</td>
<td>13.2</td>
<td>55.0</td>
<td>34.0</td>
</tr>
<tr>
<td>Read speech</td>
<td>70,223</td>
<td>16.9</td>
<td>56.9</td>
<td>13.2</td>
<td>24.1</td>
<td>67.8</td>
<td>23.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of speech</th>
<th>#words</th>
<th>% mono</th>
<th>% mono emb</th>
<th>%P+emb</th>
<th>%P 2syll</th>
<th>%P 3syll</th>
<th>%P 4syll</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retold story</td>
<td>3,502</td>
<td>37.5</td>
<td>66.0</td>
<td>3.6</td>
<td>11.2</td>
<td>81.1</td>
<td>15.4</td>
</tr>
<tr>
<td>Retold vacation</td>
<td>2,469</td>
<td>34.5</td>
<td>65.6</td>
<td>3.4</td>
<td>8.6</td>
<td>75.0</td>
<td>15.4</td>
</tr>
<tr>
<td>Informal vacation</td>
<td>4,565</td>
<td>19.5</td>
<td>67.5</td>
<td>5.2</td>
<td>16.0</td>
<td>71.6</td>
<td>24.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of speech</th>
<th>#words</th>
<th>% mono</th>
<th>% mono emb</th>
<th>%P+emb</th>
<th>%P 2syll</th>
<th>%P 3syll</th>
<th>%P 4syll</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spont. conv. (face-to-face)</td>
<td>106,182</td>
<td>7.6</td>
<td>75.3</td>
<td>5.5</td>
<td>25.1</td>
<td>75.9</td>
<td>18.9</td>
</tr>
<tr>
<td>Interv. w/Dutch teachers</td>
<td>25,687</td>
<td>11.3</td>
<td>71.0</td>
<td>5.1</td>
<td>18.2</td>
<td>71.5</td>
<td>21.5</td>
</tr>
<tr>
<td>Spont. telephone dialog.</td>
<td>201,141</td>
<td>5.1</td>
<td>75.8</td>
<td>7.3</td>
<td>37.8</td>
<td>70.9</td>
<td>23.7</td>
</tr>
<tr>
<td>Sim. business negotiations</td>
<td>25,485</td>
<td>8.4</td>
<td>72.1</td>
<td>4.8</td>
<td>17.2</td>
<td>73.4</td>
<td>18.3</td>
</tr>
<tr>
<td>Interviews/discussions (B)</td>
<td>75,106</td>
<td>10.1</td>
<td>65.5</td>
<td>7.3</td>
<td>19.5</td>
<td>68.4</td>
<td>22.0</td>
</tr>
<tr>
<td>Political interv/disc. (nB)</td>
<td>25,117</td>
<td>12.6</td>
<td>59.5</td>
<td>7.2</td>
<td>13.6</td>
<td>57.1</td>
<td>31.3</td>
</tr>
<tr>
<td>Lessons in classroom</td>
<td>25,961</td>
<td>13.3</td>
<td>67.0</td>
<td>7.9</td>
<td>23.5</td>
<td>67.0</td>
<td>25.1</td>
</tr>
<tr>
<td>Live commentaries (B)</td>
<td>24,986</td>
<td>12.1</td>
<td>64.3</td>
<td>7.0</td>
<td>16.2</td>
<td>66.1</td>
<td>23.9</td>
</tr>
<tr>
<td>News reports (B)</td>
<td>25,065</td>
<td>15.2</td>
<td>64.5</td>
<td>7.1</td>
<td>18.8</td>
<td>63.1</td>
<td>27.4</td>
</tr>
<tr>
<td>News (B)</td>
<td>25,296</td>
<td>21.6</td>
<td>49.8</td>
<td>12.4</td>
<td>16.6</td>
<td>47.4</td>
<td>34.2</td>
</tr>
<tr>
<td>Comment/columns (B)</td>
<td>25,071</td>
<td>19.0</td>
<td>61.1</td>
<td>8.0</td>
<td>17.1</td>
<td>67.0</td>
<td>21.8</td>
</tr>
<tr>
<td>Speeches/sermons</td>
<td>5,184</td>
<td>21.1</td>
<td>60.3</td>
<td>6.7</td>
<td>12.7</td>
<td>67.7</td>
<td>23.4</td>
</tr>
<tr>
<td>Lectures/seminars</td>
<td>14,913</td>
<td>17.0</td>
<td>63.0</td>
<td>5.8</td>
<td>13.2</td>
<td>55.0</td>
<td>34.0</td>
</tr>
<tr>
<td>Read speech</td>
<td>70,223</td>
<td>16.9</td>
<td>56.9</td>
<td>13.2</td>
<td>24.1</td>
<td>67.8</td>
<td>23.1</td>
</tr>
</tbody>
</table>

results. ‘Tot’ (al) shows the percentage of words of a given word length (WL), where length is denoted in terms of total number of syllables. The row ‘Emb’ shows for each word length, the percentage of polysyllabic words with lexical embedding (at any position in the word). Thus for all bisyllabic words, 70.6% had lexical embedding. Note, length ‘1’ indicates the percentage of monosyllabic words that were embedded in a polysyllabic word. Not surprisingly, the percentage of polysyllabic words that have at least one embedded word rises with increasing number of syllables. The most frequently embedded word is ‘de’ (/d@/), the. It is most often embedded word-finally, e.g., to create the past tense form of verbs.

We further analysed the lexicon with respect to word-initial embedding. Row ‘Wi’ in Table 3 shows word-initial embedding per word length; length ‘1’ denotes the percentage of monosyllabic words that occurs as word-initial embedding. Word-initial embedding occurs for 38.588 (47.9%) of all polysyllabic words. So, like found in other studies [4-6], the majority of lexical embedding occurs word-initially. Contrary to the results for embedding at all positions, word-initial embedding occurs more often in shorter polysyllabic words (see Table 3). The most frequently word-initially embedded word is ‘ge’ (/g@/), Flemish you – note that in the spoken Dutch analyses we only use the Northern Dutch part of the CGN; however, Flemish Dutch words are part of the TST lexicon). It most often occurs in two and three syllable words, e.g., the past participle of ‘maken’ (to make) is ‘ge+maak+t’ (/x@ma:kt/). Note that ‘ge’ only very rarely occurs in Northern Dutch; it thus has a very low frequency in the mental lexicon of Dutch people. The presence of ‘ge’ in the TST lexicon therefore results in an overestimation of the degree of lexical embedding.

We subsequently analysed lexical embedding in terms of POS tags. The TST lexicon provides (possibly multiple) POS words, divided over 14 different speech styles/components (see Table 1). For each word, a manually verified phonemic transcription is available, as well as a manually verified word segmentation. The IFA corpus data consist of 10,536 words (see Table 1). For each word, a manually verified phonemic transcription is available as well as a manually verified word-initial embedding; length 1 for Emb and Wi denotes the % word types that were embedded.

Table 3. TST statistics per word length in #syllables: Tot:

<table>
<thead>
<tr>
<th>Word length</th>
<th># of words</th>
<th>% of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>106,182</td>
<td>75.3%</td>
</tr>
<tr>
<td>2</td>
<td>25,687</td>
<td>71.0%</td>
</tr>
<tr>
<td>3</td>
<td>201,141</td>
<td>75.8%</td>
</tr>
<tr>
<td>4</td>
<td>75,106</td>
<td>65.5%</td>
</tr>
<tr>
<td>5</td>
<td>25,117</td>
<td>59.5%</td>
</tr>
<tr>
<td>6</td>
<td>25,961</td>
<td>67.0%</td>
</tr>
<tr>
<td>7</td>
<td>24,986</td>
<td>64.3%</td>
</tr>
<tr>
<td>8</td>
<td>25,065</td>
<td>64.5%</td>
</tr>
</tbody>
</table>

3. The analyses

3.1. TST lexicon analysis

4,613 of the 85,150 word types in the TST lexicon were monosyllabic (5.4%). 78.4% (63,180) of all polysyllabic words contained an embedded word; 35.6% of these (22,480 words) contained at least two embedded words; the average number of embedded words for all polysyllabic words with lexical embedding was 1.4. Table 3 shows more detailed
tags for each word in the lexicon. For instance, ‘waar’ has
eight different POS tags (the maximum number found for a
monosyllabic word), having different meanings depending on
the POS tag, e.g., ‘true’ (adjective) and ‘goods’ (noun).
However, not all of these meanings can occur as the first
syllable of a matrix word. For instance, the ‘waar’ in
‘waarheid’ (truth) has to be an adjective. Using all POS
combinations for the embedded and matrix words would result
in an overestimation of the results. However, for the CELEX
subset of the TST lexicon, POS tag information is available for
each constituent in a compound word; i.e., for ‘waarheid’, also
the POS tag of ‘waar’ is given. For 16,603 (26.3%) of all
polysyllabic words with embedding at any position and 15,003
(38.9%) of all polysyllabic words with word-initial
embedding, the POS information of the constituents is
available. The following analyses are based on these subsets.

Table 4 shows the most frequent POS tags for the
embedded words and the polysyllabic matrix words for both
embedding conditions. The results for word-initial embedding
and at any position are remarkably similar. The most frequent
POS tags for each word type are nouns. For embedded words,
the second and third most frequent are prepositions and verbs;
for matrix words, verbs and adjectives. The most frequent
form of embedding is nouns in nouns, e.g., ‘adres’ (address) in
‘adresboek’ (address book), which comprises 31.4% of all
possible POS tag combinations for embedding at all positions
and 33.8% for word-initial embedding. The second most
frequent is prepositions in verbs, e.g., ‘uit’ (out) in ‘uitgaan’
(to go out): 18.7% (all positions) and 20.5% (word-initial);
third most frequent are adverbs in verbs (7.0%, all positions),
e.g., ‘voor’ (before) in ‘voorzitten’ (to chair) and verbs in
nouns (11.4%, word-initial embedding), e.g., ‘zweef’ (glide) in
‘zweefduik’ (swan dive). However, in Dutch, most verb stems
like ‘zweef’) can also appear as nouns, so most of the verb in
noun embeddings are actually noun in noun embeddings.

In a final analysis, the POS tags of the embedded words
were collapsed into two classes: content words (nouns, verbs,
and adjectives) and function words (the rest), 66.1% (all
positions) and 63.2% (word-initial) of the embedded words
were content words. This seems to indicate that the problem
of lexical embedding is indeed a serious one. However, this
needs to be further investigated on real speech, as not all
words have equal frequency.

3.2. Real speech analysis
An important difference between the analyses of the TST
lexicon and the speech in the CGN and IFA is the occurrence
of pronunciation variation in real speech, whereas the TST
lexicon only lists canonical pronunciations. In order to allow
for pronunciation variation, non-canonical pronunciations of a
monosyllabic word, when encountered, were added to the
possible pronunciations of that word and subsequently
searched for as part of a longer word. So, when a
pronunciation is only encountered as part of a polysyllabic
word and not as a monosyllabic word we did not take this into
account. The reason is that only when the pronunciation of the
embedded word matches the pronunciation of the start of the
matrix word, we expect ambiguity due to lexical embedding to
arise during speech processing. It may still be the case that the
particular pronunciation may occur for monosyllabic words in
real life, but if it is not in the corpora we cannot check this.
Finally, as syllable information for these pronunciation
variants is not available, our analyses of lexical embedding in
real speech only focus on word-initial embedding.

Table 1 and 2 present general results split out for each
component of the CGN and IFA corpora, respectively.

%ratio’ shows the type/token ratio. Summing the results per
corpus; CGN consists of 675,417 word tokens, which
comprise 70,188 word types, a ratio of 10; the IFA corpus
consists of 10,536 tokens, which comprise 3,056 word types, a
ratio of 29. Speech thus is (highly) repetitive, especially in the
CGN corpus, and particularly for spontaneous speech.

The column ‘%mono’ presents the percentage of word
tokens that is a monosyllabic word. For the CGN, the
categories ‘news’, ‘read speech’, and ‘political interviews’
have the lowest proportion of monosyllabic words. This is not
surprising as longer words are more often used in written than
in spoken language, and these categories are of all CGN
components closest to written language. The percentage
monosyllabic words is highest for spontaneous speech
(conversations and telephone dialogues). This is also not
surprising as in spontaneous speech, for instance, a lot of
monosyllabic interjections are used. For the IFA corpus, the
same result was found: the most spontaneous speech style
comprised the highest percentage of monosyllabic words.

More important to our research question: it seems that the
problem of word-initial embedding is not as wide-spread as
the TST lexicon analyses suggest. The percentage of
monosyllabic word tokens that could appear word-initially
embedded for the CGN was on average only 7.5% (ranging
from 4.8% for business negotiations to 13.2% for read speech; see
‘%mono emb’) and 4.1% for the IFA corpus; this is much
lower than computed for the TST lexicon (50.8%).

Furthermore, the percentage of polysyllabic words that
have word-initial embedding is on average 19.5% for CGN
(12.7%-37.8%; speeches/sermons vs. telephone dialogues) and
11.9% for the IFA corpus (see ‘%P emb’ in Table 1 and 2). In
general, this is again much lower than in the TST lexicon
(47.9%). Word-initial embedding most often occurs in
informal speech: the percentage of polysyllabic words with
word-initial embedding is highest for the spontaneous
telephone dialogues and conversations (CGN) and for
informal vacation (IFA corpus). A further analysis showed that
only in the two spontaneous CGN components and in the
‘lessons in classrooms’, ‘ge’ occurred as monosyllabic word.
As explained in Section 3.2: ‘ge’ is very frequent as a prefix in
verbs; the presence of ‘ge’ in these three components increases
the degree of lexical embedding for these three components.
The most frequent matrix word in the two spontaneous
components is ‘gewoon’ (normal; thus with ‘ge’ embedded).
The higher percentage of lexical embedding for the telephone
dialogues component compared to the others is due to more
words starting with ‘ge’, and secondly a higher frequency of
these words.

An analysis of the word length of the polysyllabic words
with word-initial embedding showed that the vast majority of
these words consisted of only two syllables (CGN: 47.4%-
75.9%; IFA: 71.6%-81.1%; see columns ‘P 2/3/4 syl’). This is
similar to the TST lexicon where the majority of words with
embedding also were bisyllabic.

The analysis into POS tags showed that the most frequent
POS tags for the matrix words were nouns (CGN: 44.0%-
51.5%; business negotiations vs. political interviews; IFA:
50.0%-57.9%; retold story vs. retold vacation) or verbs (CGN:
43.7%-52.8%; news vs. lessons in classrooms; IFA: 50.9%-
informal vacation), the second most frequent were verbs or
nouns, respectively. Third most frequent are adjectives for
CGN. The picture is more blurred for the IFA corpus. These
results match the results for the TST lexicon (see Table 4).

For the IFA corpus, monosyllabic embedded words were
most frequently prepositions (53.1%-69.8%; retold vacation
vs. informal vacation) or nouns (46.9% retold story), while
nouns or adverbs, respectively, are second most frequent. These results are in line with the TST results. The most frequent POS tags for monosyllabic embedded words in CGN are prepositions (48.9%-91.4%; speeches/sermons vs. interviews/discussions). The second most frequent are adverbs (3.4%-25.5%; news vs. speeches/sermons), and third are nouns. These results differ from the TST lexicon results: nouns are not as often embedded in other words as one would expect on the basis of the lexicon, and secondly the higher percentage of prepositions embedded differs from the TST lexicon. As explained before, Dutch is a compounding language, and long words can easily be created by adding other words (often nouns) to it. Dictionaries contain many of these large compounds, thus increasing noun embedding, whereas these long compounds occur far less frequently in everyday speech. The higher percentage of embedded prepositions can partly be explained considering that many prepositions in Dutch can be used as a particle in verbs, such as in the earlier example ‘uit+gaan’. An analysis into the forms of embeddings indeed showed that prepositions embedded in verbs (ranging 26.8%-34.2%; speeches/sermons vs. lessons in classroom; IFA: 33.3%; informal vacation) or nouns embedded in nouns (CGN: 23.5%-33.4%; read speech vs. live commentaries; IFA: 46.9%-47.4%; retold story vs. retold vacation) are most frequent. These results, in general, match the TST lexicon results, where nouns embedded in nouns was most frequent, followed by prepositions in verbs.

Finally, since nouns, verbs, and adjectives are the most frequently occurring embedded and matrix words, the vast majority of words involved in lexical embedding in the CGN and IFA corpora are thus also content words.

4. General discussion
We investigated the degree to which lexical embedding occurs in spoken Dutch by analysing lexical embedding in a large dictionary, and more importantly in speech obtained from two corpora. Previous studies on English [4,5] and Dutch [6] showed that a majority of polysyllabic words have shorter words embedded in them, and that these words are most likely to be embedded word-initially. This result was confirmed in the analysis of the TST lexicon: 78.4% of all polysyllabic words contained at least one embedded word, and 47.9% of all polysyllabic words contained a word-initially embedded word. However, this result was not found for real speech: on average 19.5% of the polysyllabic words in CGN and 11.9% in the IFA corpus had a word-initially embedded word. The vast majority of the TST lexicon consists of words that have a very low frequency of use in everyday speech. These low frequency words tend to be longer morphologically complex words, resulting in many possible embeddings. However, these words are mainly used in written language (if at all) and not in spoken language. The CGN and IFA corpora analyses have shown that many of these low frequency and polysyllabic words do not actually occur in real speech, thus reducing the potential problem of lexical embedding. Speech processing thus appears to be simpler than one might expect on the basis of statistics computed from dictionaries (like was done in [4,6]).

So, how does the degree of lexical embedding in spoken Dutch compare to that in spoken English? According to [5], 71.1% of all polysyllabic word tokens in the MARSEC corpus contained at least one embedded word, while the percentage of word types with word-initial embedding was 50-55%. A separate analysis on the spoken Dutch data showed that 23.7% of all polysyllabic word types in CGN (range: 14.5%-35.9%) and 16.5% in the IFA corpus have word-initial embedding. The results found for spoken Dutch are thus lower than those for spoken English. This is perhaps somewhat surprising considering that Dutch is a compounding language. It might be the case that in English most syllables also occur as monosyllabic words, whereas this might be less so in Dutch; this needs further investigation. Another issue might be the speech styles: MARSEC mainly consists of news items, commentaries, sermons, and poetry; so, speech that is more or less prepared (or even read), while the spoken Dutch corpora also contain spontaneous speech styles. Read speech differs from more spontaneous speech in that it tends to contain longer words (see, e.g., the lower percentage of monosyllabic words in the read speech component of CGN), like written language. Longer words, in turn, result in more lexical embedding (see Table 3). As is clear from our analyses, the degree of lexical embedding differs between the CGN and IFA corpus, and also between the different speech styles within each corpus (see e.g., Tables 1 and 2). More research is, however, needed to explain the differences.

To conclude, these analyses show that lexical embedding is a phenomenon that occurs in spoken Dutch, and is not limited to dictionaries. As lexical embedding is most prevalent in spontaneous speech, it thus is a phenomenon that listeners have to deal with on a large scale in everyday communication. The words most often involved in lexical embedding are content words, which is most likely a result of Dutch being a compounding language. On the bright side, content words tend to be reduced less often than function words, which is helpful for speech recognition.

5. Acknowledgements
This work was undertaken while Stefanie Okoloszki was on placement in the Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands. Odette Scharenborg was supported by a Veni-grant from NWO. The authors would like to thank Eric Sanders for help with some scripts, and Lou Boves and Mirjam Ernestus for useful comments and suggestions on an earlier version of this paper.

6. References