WORDS WITHIN WORDS IN A REAL-SPEECH CORPUS

Anne Cutler, James McQueen, Harald Baayen and Hans Drexler

MPI for Psycholinguistics, Wundtlaan 1, NL-6525 XD Nijmegen.

ABSTRACT - In a 50,000-word corpus of spoken British English the occurrence of words embedded within other words is reported. Within-word embedding in this real speech sample is common, and analogous to the extent of embedding observed in the vocabulary. Imposition of a syllable boundary matching constraint reduces but by no means eliminates spurious embedding. Embedded words are most likely to overlap with the beginning of matrix words, and thus may pose serious problems for speech recognisers.

INTRODUCTION

The efficiency of spoken word recognition depends upon, among other factors, the pattern of occurrence of words within other words. Consider, for example, the word *candlestick*. This is a compound word made up of *candle* and *stick*. But *candle* itself contains *can* as its first syllable, and *can* is itself an English word. If a recognition system processes input in strictly left-to-right sequence, then the first word it encounters in *candlestick* will be *can*. The onset of a new word would then be erroneously postulated at the /d/. Since this is not in fact a new word, at the very least some backtracking is required. Beginning again from the onset of *candlestick*, the next word encountered after *can* is *canned*, which is again not in fact present; yet another attempt would produce as the next candidate (for speakers of a non-rhotic dialect) *candour*, and *candle* would appear only on the fourth attempt. If within-word embedding occurs to a considerable degree, it could in principle be causing continued problems for recognition systems.

Thus it is important, both for theories of human spoken-word recognition and for practical implementation of automatic recognition systems using realistic vocabularies, to know how large the problem of embedding actually is. Statistical analyses of actual vocabularies have been conducted to this end. For example, McQueen and Cutler (1992) analysed the vocabulary of English, using a 25,000-word phonetically transcribed lexicon based on the *Longman Dictionary of Contemporary English*. Their results showed that embedding within the English vocabulary is very widespread. Even counting only embedded words whose syllable boundaries matched those of the matrix word (thus for example, *can* in *candle* but not in *scandal*), a majority of polysyllabic words contained other words embedded within them. Moreover, embedding was most frequent in word-initial position. Just the number of monosyllabic words forming the first syllable of longer words amounted to 94% of all the analysed words (note that homophony resulted in some initial syllables contributing more than one word embedded in the same position).

An analysis of the Dutch vocabulary by Frauenfelder (1991) yielded highly comparable results. Along similar lines, Luce (1986) computed that most short words in English do not become unique until at or after their end, i.e. they could potentially be continued to become other words (*can* could become *cant*, *candle*, *canteloupe* etc.).

It is therefore clear that within the vocabulary, embedding is substantial enough to constitute a potential problem for recognition systems. Needless to say, however, vocabulary studies can only provide a very rough indication of the size of the problem in real speech. It is not the case that every word in the vocabulary has an equal chance of being encountered in real speech. It could, for instance, be the case that within-word embeddings occur primarily in rather rare long words, which are seldom heard in normal spoken language. In this case the problem of embedding may not be substantial enough to cause concern in most everyday recognition situations. Furthermore, it could be the case that other sources of information act to minimize the problems for processing caused by embedding; for instance, a majority of embedded words could be morphologically related to their matrices (i.e. could be stems within affixed matrices), in which case recognition of the embedded word would activate the same morphological family that recognition of the matrix word would. These issues were addressed in the present project, which is as far as we know the first to investigate the occurrence of words embedded within other words in a corpus of real speech.

THE MARSEC CORPUS

The Spoken English Corpus (SEC) is about 50,000 words of spoken British English, mostly taken from radio broadcasts. The corpus was collected in the 1980s, largely by the University of Lancaster and IBM UK Scientific Centre; the initial SEC was transcribed and prosodically annotated. A machine-readable version of the SEC, called MARSEC (Roach, Knowles, Varadi & Arnfield, in press), was created in 1991-3 via a collaboration between the University of Leeds and the University of Lancaster. The material exists as speech files on CD-ROM, keyed to orthographic, phonemic and prosodic transcriptions. MARSEC was the corpus used in the present study.

ANALYSES

We constructed a dictionary of all the words (7304 word types) occurring in MARSEC, with the frequency of their occurrence in this corpus (a total of 49289 tokens), and searched this dictionary for within-word occurrences of all English words listed in the CELEX lexical database (Baayen, Piepenbrock & van Rijn, 1993). As CELEX and MARSEC do not use the same phonemic transcriptions, it was further necessary to construct a mapping between these transcriptions. For both embedded and matrix words, homophones were collapsed, and all multi-word entries (such as *able-bodied seaman*) were removed from consideration. For 948 of the word types in MARSEC (13%), no corresponding entry in CELEX was available. More than 75% of these words were non-English words (e.g. *Tageblatt*), numbers (e.g. *1893*), acronyms (e.g. *PLO*), or proper names (e.g. *Leibniz*).

Two sets of analyses were carried out. In one, we counted any embedded word irrespective of syllable boundary placement (thus, *can* and *canned* and *candour* and *candle* would be found in *scandal*); this we call the "phoneme analysis". Single-phoneme words (2121 tokens, amounting to 4.3% of the corpus total of 49289 analysed tokens), were excluded from this analysis, since no embedded words could by definition be found within them. In the second analysis, we counted, following McQueen and Cutler (1992), only embedded words whose syllable boundaries matched those of the matrix word (thus, *can* would be found in *candour* and *candle* but not in *cant* or *scan* or *scandal*). From this "syllable analysis" we excluded monosyllabic words (32839 tokens, amounting to 67% of the corpus total). Syllable boundary information was extracted where possible from the CELEX database; the 948 non-CELEX words were syllabified by hand.

We also investigated the extent to which embedded words are in fact morphologically legal constituents (in the first instance, stems) of their matrix word. Here we made use of the information in CELEX concerning the inflectional and derivational properties of words. These analyses omitted the 948 non-CELEX words, most of which were in any case inappropriate for an analysis based on English morphology.

Note that none of the analyses we performed is able to yield an exact assessment of the embedding problem for a recogniser. For example, if we assume that allophonic variation due to syllable structure factors is indeed taken into account in recognition, so that, for example, *can* is effectively not present in *scan* because after /s/ the phoneme /k/ is not released as it would be in syllable-initial position, then we have also to reckon with the possibility that such information might also mislead; thus voiceless stops after /s/ may be erroneously perceived as voiced, so that, for example, although *candour* may not effectively be heard in *scandal*, *gander* might well be. Our analyses can thus merely afford an approximate estimate of the scale of the embedding problem even at the isolated-word level. Far more complicated yet is the situation in continuous speech, in which embedding is possible *across* word boundaries, so that, for instance, *candle* might in principle be found in *pecan delight*. Further analyses of the MARSEC corpus will attempt to address such more complex issues; the present analyses were restricted to within-word embeddings, and attempted first to compare the frequency of within-word embedding in a real-speech corpus with the frequencies already established for the vocabulary as a whole.

RESULTS

For the phoneme-based analysis of all words longer than just one phoneme, only 7.7% of word tokens (3643 tokens) contained no embedded words; 40.3% (19013 tokens) contained a single embedded word; and 52% (24492 tokens) contained two or more embedded words. For the syllable-based analysis of polysyllabic words, 28.9% (4741 tokens) contained no embedded words; 43.6% (7168 tokens) contained a single embedded word; and 27.5% (4521 tokens) contained two or more embedded words.

Thus only a minority of words were free from embeddings; and syllable boundary constraints, although of course they rule out the possibility of embedding within monosyllabic words, by no means remove the problem for polysyllabic words. It is therefore clear that embedding is not a problem confined to the less common portion of the English vocabulary; it occurs very frequently in this corpus of real speech.

We next consider the question of where words are embedded within their matrices - at the beginning (*can* in *cant*), middle (*can* in *scant*) or end (*can* in *scan*). Figure 1 shows, for both phoneme and syllable analyses, the frequency of occurrence in MARSEC word types of embedded words as a function of the position in the matrix word at which the embedded word occurs. In this computation we also compared frequency of embedding both with and without a morphological constraint. It can be seen that in both phoneme and syllable analyses there is a very strong tendency for the embedded word to occur early in the matrix word, i.e. to overlap from the first phoneme or syllable respectively. Excluding embedded words which are morphological stems of their matrix words (the filled circles in Figure 1) makes virtually no difference to the shape of the curve; only in the initial position is there even a noticeable difference. This difference arises as a result of excluding suffixed words such as *argument* embedded in *arguments* or *can* in *canned*. However even when these morphologically related embeddings are excluded, overlap from initial position remains far and away the modal case.

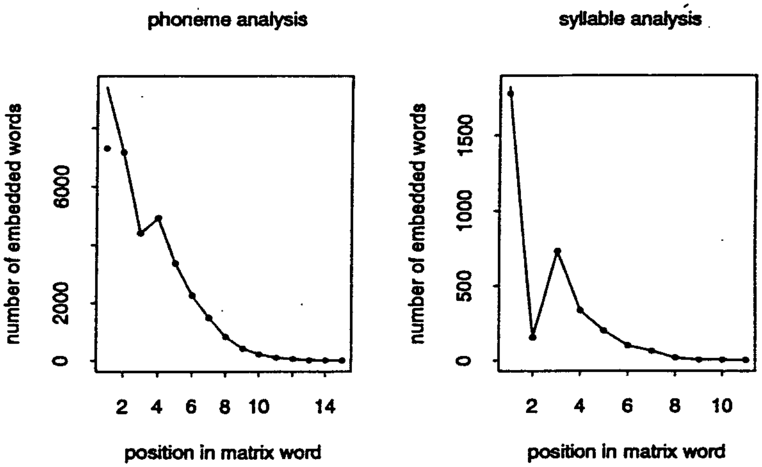

Figure 1. Number of embedded words as a function of the position in matrix word types at which the embedded word occurs for phoneme-based (left) and syllable-based (right) analyses. The analyses in which all embedded words are counted irrespective of possible morphological relationships are represented as unbroken lines. The analyses in which morphological constituents are not counted as embedded words are represented by filled circles not connected by lines.

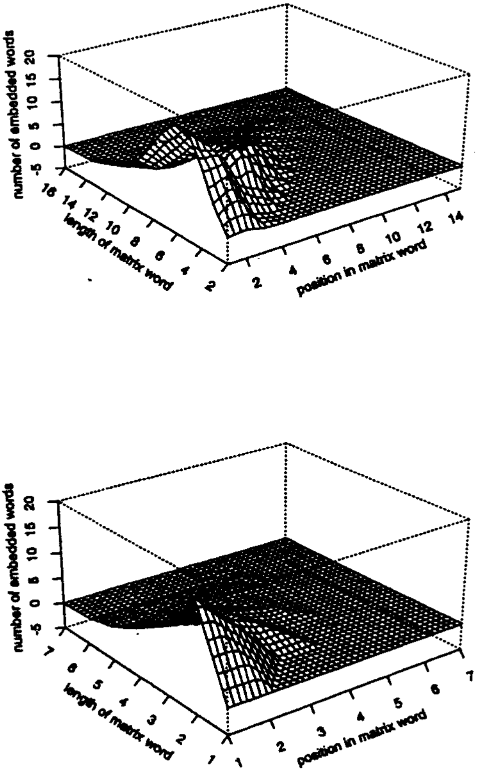

Figure 2. Number of embedded words (in units of 100) as a function of the position in the matrix word at which the embedded word occurs, and of the length of the matrix word for phoneme-based (top) and syllable-based (bottom) counts.

Naturally, the distribution of word lengths is not uniform within the corpus; there are more shorter words and fewer longer words. Thus it is reasonable to ask whether the skewed distribution of embedding position (more likely at the beginning than at the end of matrix words) might at least in part be due to asymmetries in the word length distribution. To answer this question, we computed the number of embedded words as a function simultaneously of position in the matrix word and length of matrix word. Figure 2 shows the results of this analysis in a three-dimensional representation, again for the phoneme and the syllable analysis separately. It can be seen that the trend towards embedding being most likely in the early positions is present for matrix words of all lengths independently - irrespective of matrix word length, the highest point in the position curve is always at the early positions. Thus this trend is not the trivial consequence of the fact that there are fewer possibilities of later-position embeddings, given that many shorter matrix words have fewer later positions; it is generally true for the words in this natural corpus.

Further analyses show that there is an unsurprising effect of word length on the frequency of embedding: the mean number of embeddings increases with the length of the matrix word. This function is remarkably linear - adding one more phoneme to the matrix word increases the mean number of embeddings by about 0.9. Words of two phonemes have, on average, one embedded word, while those twelve phonemes in length have, on average, ten embeddings. Even after excluding single-phoneme embedded words the number of embeddings remains high (for example, matrix words twelve phonemes in length still contain, on average, seven embedded words). Thus the high proportions of embeddings which we observe are not simply due to, for example, every occurrence of the phoneme schwa being counted as the word *a*. Although there are many such single-phoneme embeddings, words with such embeddings almost always have other embeddings.

CONCLUSION

Firstly, our analyses have revealed that within-word embedding occurs very often in real speech; it is not a phenomenon which is restricted to the obscurer portions of the English vocabulary. Secondly, embeddings detected by our computational analyses are not an artefact of morphological processes such as suffixation; excluding such related embedding-matrix pairs does not significantly reduce the frequency of embedding. Thirdly, the position at which embedded words occur within their matrices is, overwhelmingly, towards the beginning. Fourthly, information about syllable boundary occurrence reduces the extent of embedding at the word level by confining it to polysyllabic words, and this is a significant reduction simply because monosyllabic words form a large proportion of the corpus; but there is still extensive embedding, matching syllable boundaries of the matrix, within polysyllabic words.

These findings have far-reaching implications for models of human speech recognition and for application of automatic speech recognition to real speech input. Above all, the high frequency of embedding in the initial portions of matrix words means that a recogniser cannot assume that the first word it encounters in a sequential input is indeed the intended word. Furthermore, the possibility of cross-boundary embeddings in continuous speech can only heighten the problem. Strictly sequential models of recognition (such as, in psycholinguistics, those proposed by Cole and Jakimik [1978] or Marslen-Wilson and Welsh [1978]) are therefore unlikely to give an adequate account of speech recognition.

More recent models of human speech recognition (e.g. the TRACE model of McClelland and Elman [1986], or Norris' SHORTLIST [1994]) postulate a process of competition between candidate words. Competition allows the recogniser to solve the problem posed by embedding in the following way: all words compatible with the input - intended matrix words and spurious embeddings alike - will be activated, and will compete with one another, until the competition process is won by any sequence of words which (presumably uniquely) accounts for the entire input. Experimental evidence now supports the presence of active competition between candidate words in human spoken-word recognition (McQueen, Norris & Cutler, 1994; Norris, McQueen, & Cutler, in press; Vroomen & de Gelder, in press).

Moreover, competition works even more efficiently when augmented by exploitation of statistical probabilities in the vocabulary, e.g. the metrical asymmetries within English which entail that by far the majority of content words begin with a strong syllable (Cutler & Carter, 1987). (Note that the analyses of embedding within the vocabulary carried out by McQueen and Cutler [1992] distinguished between words of different metrical structures, and between the content and function word portions of the vocabulary. We have not done that in the present study, but the closeness of the present results to those of McQueen and Cutler in other respects suggest that in real speech one would find what was found in the vocabulary:

both strong-initial and weak-initial matrix words contain many embeddings, but embeddings in early position in weak-initial words are more likely to be function words.) In human speech recognition, competition processes operate in tandem with exploitation of metrical structure (McQueen, Norris & Cutler, 1994; Norris, McQueen, & Cutler, in press). We would argue that our findings, suggesting that within-word embedding is widespread in real speech input, strengthen the case for a solution of this kind to the recogniser's problem of assigning speech input correctly to the intended sequence of words.

ACKNOWLEDGEMENTS

We thank Peter Roach for making the MARSEC database available to us.

REFERENCES

Baayen, R.H., Piepenbrock, R. & van Rijn, H. (1993) *The CELEX lexical database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Cole, R.A. & Jakimik, J. (1978) *Understanding speech: How words are heard.* In G. Underwood (Ed.), Strategies of Information Processing (pp. 67-116). London: Academic Press.

Cutler, A. and Carter, D.M. (1987) *The predominance of strong initial syllables in the English vocabulary.* Computer Speech and Language, 2, 133-142.

Frauenfelder, U.H. (1991) *Lexical alignment and activation in spoken word recognition.* In J. Sundberg, L. Nord & R. Carlson (Eds.) Music, Language, Speech and Brain (pp. 294-303). London: Macmillan.

Luce, P.A. (1986) *A computational analysis of uniqueness points in auditory word recognition.* Perception & Psychophysics, 39, 155-158.

Marslen-Wilson, W.D. & Welsh, A. (1978) *Processing interactions and lexical access during word recognition in continuous speech.* Cognitive Psychology, 10, 29-63.

McClelland, J.L. & Elman, J.L. (1986) *The TRACE model of speech perception.* Cognitive Psychology, 18, 1-86.

McQueen, J.M. & Cutler, A. (1992) *Words within words: Lexical statistics and lexical access.* Proceedings of the Second International Conference on Spoken Language Processing, Banff, Canada; Vol. 1, 221-224.

McQueen, J.M., Norris, D.G. and Cutler, A. (1994) *Competition in spoken word recognition: Spotting words in other words.* Journal of Experimental Psychology: Learning, Memory and Cognition, 20, 621-638.

Norris, D.G (1994) SHORTLIST: *A connectionist model of continuous speech recognition.* Cognition, 52, 189-234.

Norris, D.G., McQueen, J.M. and Cutler, A. (in press) *Competition and segmentation in spoken word recognition.* Journal of Experimental Psychology: Learning, Memory and Cognition.

Roach, P., Knowles, G., Varadi, T. & Arnfield, S. (in press) *MARSEC: A machine-readable Spoken English Corpus.* Journal of the International Phonetics Association.

Vroomen, J. & de Gelder, B. (in press) *Metrical segmentation and lexical inhibition in spoken word recognition.* Journal of Experimental Psychology: Human Perception and Performance.