THE LEXICAL STATISTICS OF COMPETITOR ACTIVATION IN SPOKEN-WORD RECOGNITION

Anne Cutler, James M. McQueen, Maarten Jansonius and Saskia Bayerl

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

ABSTRACT: The Possible Word Constraint is a proposed mechanism whereby listeners avoid recognising words spuriously embedded in other words. It applies to words leaving a vowelless residue between their edge and the nearest known word or syllable boundary. The present study tests the usefulness of this constraint via lexical statistics of both English and Dutch. The analyses demonstrate that the constraint removes a clear majority of embedded words in speech, and thus can contribute significantly to the efficiency of human speech recognition.

LEXICAL STATISTICS AND MODELS OF SPOKEN-WORD RECOGNITION

The study of spoken-word recognition by human listeners has a relatively short history of some three decades. In that time, however, an important change has occurred whereby psychological models of language processing, which of course need to be primarily constrained by empirical data from laboratory studies of listening, have also come to be strongly influenced by computational analyses of the vocabulary and of real speech corpora. This "reality check" has led to the abandonment of some approaches in favour of others which promise better returns given the structure of speech input.

The earliest models of spoken-word recognition in continuous speech (e.g. Cole & Jakimik, 1978; Marslen-Wilson & Welsh, 1978) were based on the insight that speech - in contrast to written text, for which all previous processing models had in the first instance been constructed - is a temporal phenomenon. Thus words arrive at the listener's ear in sequence; the models embodied the assumption that recognition of any one word provided information about where recognition of the following word should commence, which in turn solved the continuity problem - i.e. the fact that words in speech abut one another without intervening intervals. However, this assumption simply did not survive the advent of on-line dictionary resources in the early 1980s, exemplified for instance by the landmark study of Luce (1986). Luce analysed a 20,000-word dictionary of English, in combination with frequency statistics, and established that a majority of words cannot be uniquely identified until at or after their ends. Particularly monosyllabic words are unlikely to be identifiable until after their end. *Star*, for instance, could continue as *start* or *starch* or *stark* or *starling*; *start* too could continue as *startle*; and so on. Thus it was impossible to maintain the assumption that reaching the end of a word in a speech input signal would automatically entail that a new word's beginning would follow, and the simplistic models which were based on this assumption fell into disuse.

The following generation of models (and indeed all current models) were based on simultaneous activation of multiple lexical candidates, and competition between these concurrently activated words. The competition proposal responds to the now well-established abundance of embedding within language vocabularies. Since vocabularies of hundreds of thousands of words are constructed from a phonemic repertoire of only a few dozen contrasting sounds, it is an inevitable consequence that words resemble one another and will often occur embedded within one another. Thus McQueen and Cutler (1992) found 63257 embedded words within a 24279-word dictionary of English two- to six-syllable words, an average of 2.6 embeddings per carrier word, whereby only embeddings with syllable boundaries matching those of the carrier word were taken into account (e.g. in *scandal*, *scan* was counted but *can* and *candle* were not). McQueen, Cutler, Briscoe and Norris (1995) found that 84% of English polysyllables have other words embedded within them (again, with syllable boundaries respected). Cutler, McQueen, Baayen and Drexler (1994) found that embedding was similarly rife within a real-speech English corpus; 92.3% of words in the corpus contained some embedded word, and 71.1% contained embedded words with syllable boundaries aligned with those of the carrier (the lower proportion than in the analyses above is due to the inclusion of monosyllabic words in the corpus analysis). By allowing words to become active if they are supported by the speech signal, and then to compete among themselves for recognition, models such as TRACE (McClelland & Elman, 1986), Shortlist (Norris, 1994) or the Neighborhood Activation Model (Luce & Pisoni, 1998) naturally cope with the embedding problem. Although a phrase like *free car scandal* (in Australian English) might activate *freak*, *cask*, *scan*, *can*, *candle,* and others, the embedded forms would be unable to muster sufficient activation to overcome the total activation of the intended words.

However, competition is not the only mechanism involved in the human recognition of words in continuous speech. Explicit segmentation procedures are also drawn on by listeners, and these too are well supported by lexical statistical evidence. Thus the Metrical Segmentation Strategy (Cutler & Norris, 1988) proposes that in English listeners treat strong syllables as likely word onsets; and indeed, lexical statistics show that this strategy will have a high success rate since most English words begin with strong syllables, and most strong syllables in English speech are word-initial (Cutler & Carter, 1987). Failure of such a procedure (such as erroneous recognition of *lease* in *police*) would moreover involve less than 1% of words in typical speech corpora (Cutler & McQueen, 1995).

THE POSSIBLE WORD CONSTRAINT

The Possible Word Constraint (PWC) has been proposed by Norris, McQueen, Cutler and Butterfield (1997) as one of the segmentation mechanisms whereby competition models constrain the activation of competitors to increase the efficiency of the recognition process. Their claim was based on laboratory evidence that listeners find it much harder to detect words abutted to contexts consisting only of consonants than in contexts containing a vowel - thus *apple* was harder to detect in *fapple* than in *vuffapple*. Norris et al. proposed that this effect arises from a constraint on activation whereby words which would leave a vowelless residue between their edge and the nearest known boundary would have their activation reduced. In this form the constraint was incorporated into the Shortlist model, producing successful simulations of the laboratory findings (Norris et al., 1997). The laboratory effect demonstrated for English replicates in other languages (McQueen & Cutler, 1998; McQueen, Otake & Cutler, 2001; Cutler, Demuth & McQueen, 2002), and is insensitive to language-specific vocabulary structure (Norris, McQueen, Cutler, Butterfield & Kearns, 2001; Cutler et al., 2002). Nor is it merely a side-effect of syllabic segmentation of speech input, since consonant-only residues and syllabic residues do not produce different response patterns in an initial-syllable segmentation task (Kearns, Norris & Cutler, 2002).

The putative explanation for the PWC is that it efficiently rules out spurious embeddings which do not preserve syllable boundaries, such as *ring* in *bring* or *zoo* in *zoom*; such words may be activated, but would be penalised on the grounds that they left a vowelless residue - [b] and [m] respectively - and as a result they would no longer mount serious competition for their carrier words. Note, incidentally, that there is independent evidence that listeners actually do make use of syllable boundary constraints in segmenting speech (McQueen, 1998).

But how useful would the PWC in fact be? Kearns et al. (2002) calculated a few statistics relevant to their study, which tested CV and CVC words embedded in longer nonwords with consonantal or syllabic residues (e.g. *zoo* in *zooth*, *zoothig* or *bell* in *belsh*, *belshig*; the former would in each case be ruled out by the PWC). The vocabulary as a whole includes more embeddings of monosyllables in initial position which leave syllable residues (*zoo* in *zulu*, *bell* in *bellows* - these would not be affected by the PWC), but predictions for actual speech taking frequency into account produce far more embeddings leaving consonantal residues (*zoo* in *zoom*, *bell* in *belt*, which would be ruled out by the PWC). This suggests that the PWC would indeed be very useful to listeners. The current study tests this prediction against the vocabulary as a whole. Note that the PWC does not simply apply to words embedded at the edge of other words (*ring* in *bring*, *zoo* in *zoom*). It also rules out internally embedded words which leave a vowelless residue between their edge and the nearest syllable or word boundary - for instance, *ring* in *trinket*, *sell* in *myself*, *eye* in *consignment*. Since the apparent universality (i.e. insensitivity to language-specific structure) of the PWC predicts that it should be useful for any language, and since both English and Dutch vocabularies were available to us, we carried out the analyses for both these languages. The effects in English which motivated Norris et al.'s (1997) PWC proposal have also been demonstrated in Dutch (McQueen & Cutler, 1998).

METHOD

The analyses were conducted on the CELEX database (Baayen, Piepenbrock & Van Rijn, 1993). The CELEX database for English comprises more than 70,000 words; the database for Dutch is substantially larger, over 280,000 words. Further, CELEX permits not only vocabulary statistics, but also estimates of likely real-speech occurrence via frequency statistics based on a corpus of 17.9 million words for English and 42.4 million words for Dutch. The primary intention of the study was to tally embedded words which would be penalised by the PWC (*can* in *scan*, *cant* or *scant*) versus those which would not (*can* in *pecan*, *canny* or *mechanic*).

For the purpose of the study we adapted CELEX in certain respects. Starting with the CELEX wordform database, entries for forms with identical phonetic notation - disregarding stress marks and syllable boundaries - were collapsed, and frequency counts summed across instances of a single phonetic form. Also, some entries were removed from the database: clitic forms (*'s*, *'d*), multi-word phrases (*worked out*, *emergency room*), letter names (*ef*), vowelless words (*ssh*) and abbreviations. The final number of evaluated forms was 71,187 for English and 281,580 for Dutch.

Each word in the corpus, in its phonetic notation without syllable or stress information, was checked against all shorter words in the corpus to find possible embeddings. For each language, this calculation took several days to run on a single Sun-Sparc computer. The resulting raw output consisted of a list of basic records, each comprising a carrier word, an embedded word, and the left and right contexts in which the embedded word was found, all in phonetic notation.

This initial list was then enriched with additional information from the main database: frequency counts for the carrier and embedded word, syllable count for each word, and exact syllabification of each word. Using the syllabification information, the position of the embedded word in the carrier could be compared with the carrier's syllable boundaries, to determine whether the embedding was aligned with the syllable structure of the carrier; if not, it could be further determined whether a vowelless residue in the carrier remained between the embedded word and the nearest syllable boundary.

Table 1. Data fields used in the embedding computations.

1. carrier orthography	14. frequency count, embedded word
2. carrier CELEX wordform ID	15. syllable count, embedded word
3. carrier CELEX lemma ID	16. embedding position (initial/medial/final)
4. carrier in phonetic notation, with syllable marks	17. alignment embedding at left with carrier syllable boundary? (Y/N/-)
5. carrier word as used in search (phonetic notation only)	18. non-syllabic residue left of embedding, if present (eg: C, CC)
6. frequency count, carrier	19. syllable count left of embedding (0 for initial embedding)
7. syllable count, carrier	20. context left of embedding, as found
8. embedded word orthography	21. context left of embedding, with syllable marks from carrier
9. embedded word CELEX wordform ID	22. alignment embedding at right with carrier syllable boundary? (Y/N/-)
10. embedded word CELEX lemma ID	23. non-syllabic residue right of embedding, if present (eg: C, CC)
11. embedded word in phonetic notation, with syllable marks	24. syllable count right of embedding (0 for final embedding)
12. embedded word as found in search	25. context right of embedding, as found
13. embedded word, with syllable marks from carrier inserted	26. context right of embedding, with syllable marks from carrier

For each case of embedding, a final data record was created with the fields listed in Table 1. The structure of this enriched listing was identical for the Dutch and the English version. All statistics reported below were computed from this listing, by summing and tallying data fields under various conditions on the information in other fields. For the PWC efficiency calculation, for instance, the most important fields are 17 and 22, from which the alignment of embeddings with syllable boundaries in the carrier word can be ascertained, and 18 and 23, which indicate a vowelless residue between the embedding and the nearest syllable boundary. Using additional fields such as 16 (embedding position) or 15 (number of syllables of the embedded word), the embedding statistics could be separately computed for each embedding position (initial/medial/final) and for embedded words with differing number of syllables. Inflected forms could be identified by comparing lemma ID numbers of the carrier and the embedded word (fields 3 and 10); if these are the same, the wordforms are different inflectional forms of the same base (*can*, *canned*). Finally, the vocabulary statistics (type frequencies) could be increased by the frequency of occurrence of the relevant carrier word (field 6) to estimate likely occurrence of each type of embedding in real-speech corpora (token frequencies).

RESULTS

Overall

Table 2. Embedding frequencies for English and Dutch.

	English	Dutch
no embedded words	1088	1724
1 embedded word	4381	2164
2 or more embedded words	65718	277692
Total	71187	281580

Table 2 shows the frequency with which embedding occurs across word types in English and Dutch. In agreement with earlier analyses of English (McQueen & Cutler, 1992; Cutler et al., 1994) and Dutch (Frauenfelder, 1991), these statistics show that only a very small proportion of words (less than 2% in English and less than 1% in Dutch) contain no other words embedded within them. Table 3 shows the total number of words embedded within other words in the vocabulary, overall and separately by position (initial, medial, final), as well as token frequencies calculated from carrier word frequencies. Table 3 also breaks the embedded words down into two categories: PWC violations (-V) leaving vowelless residue (*can* in *scant* etc.) vs. non-violations (+V) leaving syllabic residue, i.e. residue containing a vowel (*can* in *mechanic* etc.).

Table 3. Type and token frequencies of embeddings which violate vs. do not violate the PWC.

	English				Dutch			
	initial	medial	final	total	initial	medial	final	total
TYPES								
-V	30896	161211	33752	225859	147811	1150273	201339	1499423
+V	72957	35857	35656	144470	462263	418907	287967	1169137
total	103853	197068	69408	370329	610074	1569180	489306	2668560
TOKENS								
-V	4172839	9748670	4886605	18808114	5371869	34157761	12793793	52323423
+V	4342280	1082056	1554770	6979106	16510920	7648582	9779434	33938936
total	8515119	10830726	6441375	25787220	21882789	41806343	22573227	86262359

For English, as can be seen, rather more than three-fifths of all embedded words violate the PWC: the ratio is 1.56 PWC violations to every one embedding which is not penalised. Taking frequency into account, however, reveals greater asymmetry: 2.69 PWC violations to every unpenalised embedding. The saving is largest for medially embedded words (e.g. *can* in *scandal*; 9.01:1) and stronger for final embeddings (*candle* in *scandal*; 3.14:1) than for initial (*can* in *cant*; approximately 1:1). For Dutch, the CELEX database is much larger. The size difference between these two CELEX word lists lies mainly in the proliferation of compounded words in Dutch, and this is of course directly reflected in the embedding statistics. A compound like *can opener* is considered two words in the English corpus, but one (*blikopener*) in Dutch. This predicts that a higher proportion of embeddings in Dutch would pass the PWC, since in compounds each word (e.g. *blik* and *opener* in *blikopener*) would strand an entire word. However, in Dutch too there are more embeddings which violate the PWC than not (1.28:1), and the asymmetry is greater when frequency is taken into account (1.54:1); again, the asymmetry is largest for medial embeddings (4.47:1) and larger for final embeddings (1.31:1) than for initial embeddings, for which the asymmetry for Dutch is in fact significantly reversed.

Table 4. Percentage of embedded words of different lengths in syllables, for each embedding position.

	English				Dutch			
no. sylls:	1	2	3	> 3	1	2	3	> 3
initial	64.27	22.45	9.30	3.78	46.92	27.44	15.30	10.34
medial	91.43	7.88	0.59	0.10	83.82	13.94	1.81	0.04
final	74.92	20.17	3.52	1.39	53.96	32.72	9.79	3.53
total	80.72	14.27	3.58	1.43	69.91	20.47	6.36	3.26

Table 4 presents some summary statistics on the type frequencies of embedded words when calculated separately by length of the embedding in syllables. For all four columns for each language in Table 4, the ratio of PWC violations to non-violations was largest for medial embeddings. It can be seen that the majority of embedded forms in both languages are monosyllabic, and that there are very few embedded long words, as would be expected. However, there is a striking difference in the patterns for words embedded initially in their carrier, medially, or finally. Briefly, embedded words tend to be longer among initial embeddings (only 64% monosyllabic in English, 47% in Dutch) and shorter among medial embeddings (91% monosyllabic in English, 84% in Dutch). This suggests that initial embeddings may include inflected forms. Inflected forms (e.g. *cans*) would often be counted as PWC violations; but one might argue that recognition of an inflected form may require recognition of the uninflected embedding, and for this reason it is desirable to tally these cases separately.

Inflected Forms

Inflected forms were identified via lemma ID comparisons as described above. These comparisons showed that dropping all embedded forms with the same lemma ID as the carrier would result in removal of 3% of the total number of English embeddings and 5% of the Dutch embeddings. In English, 100% of such cases were initial embeddings. In Dutch, 90% were initial embeddings, 3% were medial embeddings, and 7% were final embeddings. The Dutch inflected forms included past participles (prefixed or prefixed and suffixed) as well as suffixed verbs and nouns.

Syllable structure

Table 5 breaks the PWC violation cases of Table 3 down as a function of whether the vowelless residue is found between the embedding and the edge of the carrier (+C and C+ cases; e.g. *lie* in *fly*, *like*) or between the embedding and the nearest carrier-internal syllable boundary (syl+C and C+syl cases; e.g. *lie* in *apply*, *lightning*). It can be seen that the former cases predominate in English, while the compound words of Dutch lead to more of the latter cases.

Table 5. PWC violations at carrier word edge vs. internal syllable boundary.

	English		Dutch	
Context	types	tokens	Types	tokens
initial+C	24554	3790625	66762	3746916
initial+C+syl	6342	382214	81049	1624953
C+final	7643	3650509	11225	3669396
syl+C+final	26109	1236096	190114	9124397
medial+C	60405	5263983	292066	11824864
medial+C+syl	12420	462442	229896	4075273
C+medial	61095	5256737	303839	14765947
syl+C+medial	76423	3533640	633042	14068928

CONCLUSION

The analyses clearly demonstrate that the efficiency of the proposed PWC is high; its application would remove a clear majority of spuriously embedded words in speech. In English, the total saving is 73% of all such embeddings, in Dutch it is less, at 61%. The default Dutch plural inflection contains a vowel (*blikken*, 'cans'), the English equivalent does not; the Dutch diminutive also has a vowel (*blikje*). However, inflections do not comprise a large proportion of embeddings. Rather, we explain the difference as reflecting the criteria for lexical inclusion in Dutch: *can opener*, *swimming pool*, *house music* are unitary lexical entries in the Dutch database but not in the English database. It is arguably the case, though, that compounds need to count as unitary units in English too (whereby *house* may be regarded as a spurious embedding in *house music*). In that case, the English total saving may be regarded as an overestimate rather than the Dutch saving being an underestimate. Nevertheless, it appears that the PWC would be of considerable value for segmentation in both languages.

ACKNOWLEDGEMENTS

Thanks to Dennis Norris for discussion, and to Tau van Dijck for advice on large-scale calculation.

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, 67-116. London: Academic Press.

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

Cutler, A., Demuth, K. & McQueen, J.M. (2002). *Universality versus language-specificity in listening to running speech*. Psychological Science 13, 258-262.

Cutler, A. & McQueen, J.M. (1995). *The recognition of lexical units in speech*, in B. de Gelder & J. Morais (eds), Speech and Reading: A Comparative Approach, 33-47. Hove: Erlbaum.

Cutler, A., McQueen, J.M., Baayen, H. & Drexler, H. (1994). *Words within words in a real-speech corpus*. Proceedings of the 5th Australian International Conference on Speech Science and Technology, Perth; 362-367.

Cutler, A. & Norris, D. (1988). *The role of strong syllables in segmentation for lexical access*. Journal of Experimental Psychology: Human Perception & Performance 14, 113-121.

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, 294-303. Wenner-Gren International Symposium series. London: Macmillan.

Kearns, R.K., Norris, D. & Cutler, A. (2002). *Syllable processing in English*. Proceedings of the 7th International Conference on Spoken Language Processing, Denver.

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

Luce, P.A. & Pisoni, D.B. (1998). *Recognizing spoken words: The neighborhood act ivation model*. Ear and Hearing 19, 1-36.

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. (1998). *Segmentation of continuous speech using phonotactics*. Journal of Memory and Language 39, 21-46.

McQueen, J.M. & Cutler, A. (1992). *Words within words: Lexical statistics and lexical access*. Proceedings of the 2nd International Conference on Spoken Language Processing, Banff; 221-224.

McQueen, J.M. & Cutler, A. (1998). *Spotting (different kinds of) words in (different kinds of) context*. Proceedings of the 5th International Conference on Spoken Language Processing, Sydney; 2791-2794.

McQueen, J.M., Cutler, A., Briscoe, T. & Norris, D. (1995). *Models of continuous speech recognition and the contents of the vocabulary*. Language and Cognitive Processes 10, 309-331.

McQueen, J.M., Otake, T. & Cutler, A. (2001). Rhythmic cues and possible-word constraints in Japanese speech segmentation. Journal of Memory and Language, 45, 103-132.

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

Norris, D., McQueen, J.M., Cutler, A. & Butterfield, S. (1997). *The possible-word constraint in the segmentation of continuous speech*. Cognitive Psychology 34, 191-243.

Norris, D., McQueen, J.M., Cutler, A., Butterfield, S. & Kearns, R. (2001). *Language-universal constraints on speech segmentation*. Language and Cognitive Processes 16, 637-660.