WORDS WITHIN WORDS: LEXICAL STATISTICS AND LEXICAL ACCESS

James M. McQueen and Anne Cutler

MRC Applied Psychology Unit,
15 Chaucer Rd., Cambridge, CB2 2EF, U.K.

ABSTRACT

This paper presents lexical statistics on the pattern of occurrence of words embedded in other words. We report the results of an analysis of 25000 words, varying in length from two to six syllables, extracted from a phonetically-coded English dictionary *(The Longman Dictionary of Contemporary English).* Each syllable, and each string of syllables within each word was checked against the dictionary. Two analyses are presented: the first used a complete list of polysyllables, with look-up on the entire dictionary; the second used a sublist of content words, counting only embedded words which were themselves content words. The results have important implications for models of human speech recognition. The efficiency of these models depends, in different ways, on the number and location of words within words.

1. INTRODUCTION

Spoken word recognition is highly likely to be dependent on the structure of the vocabulary of the input language. The process of lexical access is one of mapping a continuous input, composed from a relatively small set of speech sounds (less than 40 phonemes in most languages [1]), onto a lexicon of tens of thousands of discrete words. This process would undoubtedly be more efficient if it took advantage of structural regularities in the vocabulary. Indeed, most models of speech recognition incorporate aspects of vocabulary structure into their operation. In particular, they are all dependent, in different ways, on the number and location of words within words.

In the Cohort Model [2,3], for instance, recognition of a word depends on how many other words share its initial portions. A word which overlaps with few other words (e.g. *choice* becomes unique on the vowel) will be recognised more quickly than a word which overlaps with many other words (e.g. *pick* becomes unique only at its offset). In the Neighbourhood Activation Model [4] however, recognition of a word depends on how many other words resemble it at any point. For example, recognition of *lice* will be affected by the fact that the words *lie*, *eye* and *ice* are all contained within it. Some models (e.g. TRACE [5]) incorporate a process of lexical competition, whereby a number of candidate words compete for portions of the input string; the number of words within words will partially determine the number of competitors, and hence will affect the recognition process. Words within words are perhaps of even greater importance for models of continuous speech recognition which incorporate explicit prelexical segmentation procedures. The Metrical Segmentation Strategy [6,7], for instance, proposes that listeners attempt lexical access at strong syllables. The efficiency of such a process depends on the number of unrelated words beginning from the first strong syllable of words with weak initial syllables (e.g. *million* in *vermilion;* [8]).

Although these models are clearly dependent on the number and location of words within words, the necessary statistics of vocabulary structure have not previously been available. In this paper we present the results of an analysis of about 25000 polysyllabic words, where we computed the frequency of occurrence of words within words.

2. ANALYSES

These analyses were based on a machine-readable version of *The Longman Dictionary of Contemporary English* [9,10]. Phonological and grammatical-class access procedures, written in Lisp, have previously been developed for this database [11,12,13]. The access code was modified, such that the word-within-word analyses could be performed in two stages. First, the database was exhaustively searched for polysyllabic words, and the phonological strings associated with each headword were extracted. These were listed in separate files, by number of syllables. These files were subdivided according to the stress of the first syllable. Thus, for words of each length, there were three groupings: words with primary stress on the first syllable, those with secondary stress, and those with an unstressed first syllable. The dictionary contains many multi-word or phrasal headwords (e.g. *hairpin bend*, *funny peculiar).* These were excluded from all searches, since their component words almost invariably have separate entries.

The second stage involved searching these base-lists for words within words. The search was based on syllable-level matches: a word was considered to be an embedded word only if it perfectly matched the syllabification of the embedding word. Thus, matches such as *mess* in *domestic* were counted, while phonemic-level matches which comprised either more or less than a whole syllable (e.g. *ten* in *stench)* were ignored.

Syllable boundaries were defined by the syllable parser built into the phonological access system [12].

This parser operates on the phonotactic constraints in Gimson [14], and the *Maximal Onset Principle* [15]. All possible locations of embedding words were searched, for embedded words up to one less syllable than the number in the embedding words. The number of embedded words of each type and at each location were counted, and each word found was listed. Several classes of dictionary headwords were excluded from the count: prefixes, suffixes, letters of the alphabet, combining forms (e.g. *-latry)* and apostrophised forms (e.g. *'re).*

We carried out two analyses. In the first, we searched **all words** in the dictionary between two and six syllables. Phrasal headwords were excluded from the base-lists, and the search was constrained as outlined above.

The second analysis was based on **content** words only. Words were only included in the base-lists if they were marked in the dictionary as being a content word (i.e. a noun, verb, adjective or adverb). Again, only non-phrasal two- to six-syllable words were included. The search was similarly constrained, so that only embedded words which were themselves content words were counted. The copular verb *be* (and its inflections), and the auxiliary verbs, being function words, were not counted in the search.

3. RESULTS

The number of words included in the base-lists for each search are given in Table 1. The results for both searches are in Tables 2 to 6. For each word length, the proportion of embedded words are listed for each possible location within the embedding words. For example, for three-syllable embedding words, embedded words could occur in the first syllable, the first and second syllables, the second syllable, the second and third syllables, or the third syllable. In each table these locations appear in sequential order, moving from top to bottom. The statistics are given in summary form, and separately according to whether the first syllables of the embedding words were unstressed or stressed (collapsing across primary and secondary stress). The data are presented as proportions: the percentage of

TABLE 1
Number of words searched

Number of syllables	Search	Initial syllable Weak	Initial syllable Strong	Sum
2	All words	1754	9848	11602
	Content	1708	9762	11470
3	All words	2689	5126	7815
	Content	2668	5099	7767
4	All words	1594	2027	3621
	Content	1592	2015	3607
5	All words	382	687	1069
	Content	382	684	1066
6	All words	55	117	172
	Content	55	117	172

TABLE 2
Proportion of words found in 2 syllable words

Word found in:	Search	1st syllable of embedding word Weak	1st syllable of embedding word Strong	Sum
1st syllable	All words	113.3%	100.4%	102.4%
	Content	44.9%	92.6%	85.5%
		unique	*canvas*	
2nd syllable	All words	84.6%	79.8%	80.5%
	Content	75.1%	49.2%	53.1%
		police	*concourse*	

TABLE 3
Proportion of words found in 3 syllable words

Word found in:	Search	1st syllable of embedding word Weak	1st syllable of embedding word Strong	Sum
1st syllable	All words	86.8%	89.9%	88.8%
	Content	43.5%	76.6%	65.3%
		bacteria	*lullaby*	
lst& 2nd syllables	All words	9.6%	42.5%	31.2%
	Content	9.2%	41.2%	30.2%
		apartment	*orthodox*	
2nd syllable	All words	83.5%	104.6%	97.3%
	Content	77.1%	29.2%	45.7%
		abandon	*microfiche*	
2nd& 3rd syllables	All words	22.5%	15.8%	18.1%
	Content	22.0%	15.5%	17.8%
		detractor	*canticle*	
3rd syllable	All words	56.8%	63.6%	61.2%
	Content	26.5%	50.6%	42.3%
		acetic	*acolyte*	

TABLE 4
Proportion of words found in 4 syllable words

Word found in:	Search	1st syllable of embedding word Weak	1st syllable of embedding word Strong	Sum
1st syll.	All words	94.6%	73.3%	82.7%
	Content	46.3%	67.9%	58.4%
		brutality	*cannibalism*	
lst& 2nd syll.'s	All words	3.6%	26.2%	16.2%
	Content	3.3%	25.9%	15.9%
		circumference	*detonate*	
1st-3rd syll.'s	All words	12.6%	7.2%	9.6%
	Content	12.5%	7.1%	9.5%
		commercialize	*operative*	
2nd syll.	All words	54.2%	94.0%	76.5%
	Content	49.0%	21.9%	33.9%
		inveterate	*photocopy*	
2nd& 3rd syll.'s	All words	21.1%	5.3%	12.3%
	Content	20.4%	5.3%	11.9%
		ionosphere	*arbitration*	
2nd-4th syll.'s	All words	11.9%	8.5%	10.0%
	Content	11.6%	8.5%	9.9%
		provocative	*absolution*	
3rd syll.	All words	99.7%	82.9%	90.3%
	Content	14.5%	73.3%	47.3%
		exclusively	*appetizer*	
3rd& 4th syll.'s	All words	6.8%	22.7%	15.7%
	Content	6.5%	22.2%	15.3%
		intensity	*undergarment*	
4th syll.	All words	35.1%	38.8%	37.1%
	Content	28.9%	21.7%	24.9%
		invalidate	*manicurist*	

words found of each length beginning in each location given the number of words searched. Examples of words in which there is an embedded word are listed in italics (e.g. in Table 2, *unique* has the word *ewe* as its first syllable, and *concourse* has the word *course* as its second syllable).

Three points about these results are worth noting. First, the percentages are those for the total number of words found over the total number of words searched. Thus. 50% would not necessarily mean that there was an embedded word in that position in half of the words searched, since in many cases more than one word was found in a given position in one word (e.g. *awe, oar* and *ore* in *audacious'*, note also that the search is based on standard British English pronunciation). 50% would thus indicate that no more than half of the words of that type contain words in that location.

TABLE 5
Proportion of words found
in 5 syllable words

Words found in:	Search	1st syllable of embedding word: Weak	Strong	Sum
1st syll.	All words	96.3%	85.6%	89.4%
	Content	39.8%	77.5%	64.0%
		bimetallism	*rationalism*	
lst& 2nd syll.'s	All words	4.2%	24.2%	17.0%
	Content	3.1%	21.6%	15.0%
		determination	*ordinarily*	
1st-3rd syll.'s	All words	10.2%	3.8%	6.1%
	Content	10.2%	3.7%	6.0%
		dishonourable	*radicalism*	
1st-4th svll.'s	All words	7.1%	5.4%	6.0%
	Content	7.1%	5.4%	6.0%
		electioneering	*haberdashery*	
2nd syll.	All words	58.6%	88.8%	78.0%
	Content	53.4%	27.8%	37.0%
		emancipation	*chemotherapy*	
2nd& 3rd syll.'s	All words	18.1%	2.9%	8.3%
	Content	17.0%	2.5%	7.7%
		emotionally	*conservationist*	
2nd-4th syll.'s	All words	5.2%	2.8%	3.6%
	Content	5.0%	2.8%	3.6%
		indefinitely	*inconsiderate*	
2nd-5th	All words	13.4%	8.7%	10.4%
	Content	13.1%	8.8%	10.3%
		inseparable	*incredulity*	
3rd syll.	All words	100.8%	57.8%	73.2%
	Content	11.3%	49.6%	35.8%
		unanswerable	*infidelity*	
3rd& 4th syll.'s	All words	37.4%	55.2%	48.8%
	Content	33.8%	53.5%	46.4%
		discriminatory	*instability*	
3rd-5th syll.'s	All words	5.0%	14.8%	11.3%
	Content	5.0%	14.9%	11.4%
		collaboration	*insupportable*	
4th syll.	All words	101.0%	97.4%	98.7%
	Content	81.4%	36.0%	52.3%
		confederation	*abracadabra*	
4th& 5th	All words	15.7%	9.6%	11.8%
	Content	14.9%	9.5%	11.4%
		contamination	*holidaymaker*	
5th syll.	All words	19.1%	20.8%	20.2%
	Content	17.3%	15.4%	16.0%
		imperialistic	*antihistamine*	

TABLE 6
Proportion of words found
in 6 syllable words

Word found in:	Search	1st syllable of embedding word: Weak	Strong	Sum
1st syll.	All words	81.8%	100.0%	94.2%
	Content	50.9%	89.7%	77.3%
		identification	*fundamentalism*	
lst& 2nd syll.'s	All words	1.8%	32.5%	22.7%
	Content	0.0%	27.4%	18.6%
			extracurricular	
1st-3rd syll.'s	All words	10.9%	*1.1%*	8.7%
	Content	10.9%	*1.1%*	8.7%
		professionalism	*radiolocation*	
1st-4th svll.'s	All words	0.0%	8.5%	5.8%
	Content	0.0%	8.5%	5.8%
			constitutionally	
1st-5th syll.'s	All words	3.6%	0.9%	1.7%
	Content	3.6%	0.9%	1.7%
		denominational	*representational*	
2nd syll.	All words	58.2%	80.3%	73.3%
	Content	56.4%	35.0%	41.9%
		emotionalism	*trinitrotoluene*	
2nd& 3rd syll.'s	All words	25.5%	1.7%	9.3%
	Content	20.0%	1.7%	7.6%
		irreconcilable	*unparliamentary*	
2nd-4th syll.'s	AH words	0.0%	4.3%	2.9%
	Content	0.0%	4.3%	2.9%
			irrecoverable	
2nd-5th syll.'s	All words	1.8%	1.7%	1.7%
	Content	1.8%	1.7%	1.7%
		denominational	*noncontributory*	
2nd-6th svll.'s	All words	9.1%	9.4%	9.3%
	Content	9.1%	9.4%	9.3%
		improbability	*nonproliferation*	
3rd syll.	All words	85.5%	100.9%	95.9%
	Content	18.2%	72.6%	55.2%
		insensibility	*onomatopoeia*	
3rd& 4th syll.'s	All words	32.7%	42.7%	39.5%
	Content	32.7%	42.7%	39.5%
		discriminatory	*reconciliation*	
3rd-5th syll.'s	All words	1.8%	4.3%	3.5%
	Content	1.8%	4.3%	3.5%
		collaborationist	*microelectronics*	
3rd-6th syll.'s	All words	*13%*	24.8%	19.2%
	Content	7.3%	24.8%	19.2%
		compatibility	*palaeontology*	
4th syll.	All words	52.7%	88.9%	77.3%
	Content	38.2%	34.2%	35.5%
		congratulatory	*plenipotentiary*	
4th& 5th syll.'s	All words	9.1%	8.5%	8.7%
	Content	7.3%	8.5%	8.1%
		elucidatory	*contraindication*	
4th-6th syll.'s	All words	0.0%	8.5%	5.8%
	Content	0.0%	8.5%	5.8%
			megalomaniac	
5th syll.	All words	56.4%	*87.2%*	77.3%
	Content	45.5%	43.6%	44.2%
		hallucinogenic	*overpopulated*	
5th& 6th syll.'s	All words	7.3%	6.8%	7.0%
	Content	7.3%	6.8%	7.0%
		ecclesiastical	*anlilogarithm*	
6th syll.	All words	18.2%	23.1%	21.5%
	Content	3.6%	18.8%	14.0%
		electrocardiogram	*incommunicado*	

Second, summing over stress patterns may obscure interesting asymmetries. For example, in 3 syllable words, although overall up to 30% contain words in the first two syllables, this is largely due to words within words beginning with strong syllables (e.g. *author* in *orthodox).* Less than 10% of weak-initial 3 syllable words have 2 syllable words embedded at their onsets (e.g. *apart* in *apartment).* Similar asymmetries are present for 2 syllable embedded words in the same position in 4 to 6 syllable embedding words. One reason for this pattern (and for others like it) is that the chance of finding an embedded word of a given type is dependent on the number of words of that type in the language. In this case, since there are many more disyllables beginning with strong than with weak syllables (see Table 1), there are likely to be more embedded words of the former than of the latter type.

Third, note that the differences between the number of words found in the content and overall searches are largest for short words with weak syllables, and smallest for long words. This is because most function words are monosyllables which can be realized as weak syllables.

4. DISCUSSION

An example will demonstrate the way in which these results can be used to assess the efficiency of different models of spoken word recognition. Consider the Metrical Segmentation Strategy (MSS; [6,7)). If lexical access is attempted at strong syllables, then false alarm recognitions may occur when embedding words contain words beginning from the first strong syllable (such as *lease* in *police* and *tractor* in *detractor).* In the content-word search, there were 2112 words of this type found in 6405 weak-initial words (i.e. 33%). Half of the words, however, are related to the words that they are within (e.g. *separable* in *inseparable;* [8]). Since we can estimate that weak-initial content words only make up about 4% of conversational English [6], we can further estimate that less than 1% (16.5% of ,4%) of words normally encountered will be weak-initial with unrelated words beginning from their second syllable. The MSS thus seems well-suited to this aspect of vocabulary structure.

The number of words found in word-initial positions highlights the general problem of embedded words for lexical access. Collapsing across all five overall searches, there were 22928 monosyllables found in word-initial position in the 24279 words searched (i.e. 94%). Thus we can estimate that polysyllabic words usually begin with other words. (More than a fifth of these embedded words were function words, since 17800 words, 74%, were found in the same position in the content-only search.) It follows that the recognition of longer words will normally involve the rejection of other, shorter words. These embedded words are perfectly consistent with the input, and appear in the more salient, word-initial position. Models of spoken word recognition therefore cannot ignore this aspect of vocabulary structure, and must provide mechanisms to deal with words within words.

ACKNOWLEDGEMENTS

This work was supported by a Joint Councils Cognitive Science/HCI Initiative Grant: E304/148. We would like to thank the Longman Group UK Inc. for allowing us to use a machine-readable version of the Longman Dictionary of Contemporary English.

REFERENCES

I. Maddieson. *Patterns of Sounds.* Cambridge-C.U.P., 1984.

2 W.D. Marslen-Wilson and A. Welsh. "Processing Interactions during Word-recognition in Continuous Speech," *Cognitive Psychology,* vol. 10, pp. 29-63, 1978.

3 W.D. Marslen-Wilson. "Functional Parallelism in Spoken Word-recognition," *Cognition,* vol. 25, pp. 71-102, 1987.

4 S.D. Goldinger, P.A. Luce, and D.B. Pisoni. "Priming Lexical Neighbors of Spoken Words: Effects of Competition and Inhibition," *Journal of Memory and Language,* vol. 28, pp. 501-518, 1989.

5 J.L. McClelland and J.L. Elman. "The TRACE Model of Speech Perception," *Cognitive Psychology,* vol. 18, pp. 1-86, 1986.

6 A. Cutler and D. Carter. "The Predominance of Strong Initial Syllables in the English Vocabulary," *Computer Speech and Language,* vol. 2, pp. 133-142, 1987.

7 A. Cutler and D. Norris. "The Role of Strong Syllables in Segmentation for Lexical Access," *Journal of Experimental Psychology: Human Perception & Performance,* vol. 14, pp. 113-121, 1988.

8 A. Cutler and J.M. McQueen. "The Recognition of Lexical Units in Speech." In B. de Gelder and J. Morais (Eds.), *From Spoken to Written Language.* Cambridge, MA: MIT Press, in press.

9 P. Proctor (Ed.). *Longman Dictionary of Contemporary English.* London: Longman, 1975.

10 B. Boguraev and E.J. Briscoe (Eds.). *Computational Lexicography for Natural Language Processing.* London: Longman, 1989.

11 J. Carroll. *The Lexical Database User Manual.* Unpublished manuscript, Computer Laboratory, University of Cambridge, 1991.

12 D. Carter. "LDOCE and Speech Recognition." In B. Boguraev and E.J. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing,* pp. 135-152. London: Longman, 1989.

13 J.M. McQueen and E.J. Briscoe. "A Computational Tool for Examining Lexical Segmentation in Continuous Speech." *Proceedings of EUROSPEECH '91,* vol. 2, pp. 697-700.

14 A. Gimson. *An Introduction to the Pronunciation of English, 3rd Edition.* London: Edward Arnold, 1980.

15 E.0. Selkirk. *Phonology and Syntax: The Relation between Sound and Structure.* Cambridge, MA: MIT Press, 1984.