Vincent van Heuven & Peter Hagman

Lexical statistics and spoken word recognition in Dutch

1. Introduction

Is the word onset special?

Spoken and visual word recognition differ crucially in that information during speech enters the sensory system sequentially (from early to late, or from left to right), whereas graphic information is made available in parallel.

It is by no means easy to see how the listener is able to recognize words in the stream of sounds that enter his auditory system. We know that an accurate and detailed image of the actual speech sounds is available to the listener only for some 100 ms. This information decays rapidly from auditory memory, and is generally lost within 250 ms after the original stimulation. Given that the majority of the words in languages such as Dutch or English take up more that 250 ms (roughly the duration of one syllable), the human word recognition system cannot afford delaying decisions until all the acoustic information pertaining to the word's identity has been heard, but must act on the incoming information as long as it is available, and recode this information into some higher-order code that is more resistant to decay over time. There are indeed strong indications that during normal, fluent word recognition in connected speech (so called 'on-line' word recognition) not only monosyllabic words, but also longer, polysyllabic words are recognized at roughly 200 ms after the word onset (Marslen-Wilson, 1985).

Given that speech is primary and writing secondary, one would predict that languages should have evolved such that the word onset carries more information as to the word's identity than the later portions of the word. It is the purpose of this paper to explore the question if indeed the distribution of information over the word forms in the (Dutch) lexicon is skewed and biassed towards the beginning of words, from a statistical point of view. We are not concerned here with the testing of a human word recognition model; we are only interested in checking specific distributional properties of the lexicon, which should logically follow from the combined effects of echoic memory limitations and the sequential nature of spoken words.

Approach: examining segmental and prosodic information

Words differ primarily in terms of their segmental structure: the specific sequence of consonants and vowels. We shall try to answer the question raised above by examining the distribution of phonemic contrasts over the word forms in a large computer-accessible Dutch lexicon, in several different ways, which we shall not outline here, but which will be described in our analysis and results section.

Moreover, the words in the Dutch lexicon do not merely contrast segmentally but also prosodically, e.g., in terms of number of syllables and stress position. It is unclear at this moment to what extent prosodic characteristics of words contribute to spoken word recognition in languages with lexical stress, such as Dutch and English. An extreme position is taken by Cutler

(1987), who explicitly denies that information on the stress pattern of a word helps to narrow down the set of alternatives from which the word will eventually be selected. In her view, stress information comes available only after the word has been accessed in the mental lexicon. An alternative view would be that prosodic information (especially stress position) may indeed help to limit the search space in the mental lexicon as the word develops in time, thus speeding up the process of lexical acces. Since prosodic information such as word length (i.e. number of syllables) and stress position may well be important to word recognition, we decided to include these factors in our statistics along with segmental information.

2. The lexical database

In order to explore the distribution of segmental and prosodic information over the words in the language we need a computer-accessible Dutch lexicon with a phonemic code specifying per word the identity of its phonemic segments, as well as the position of syllable boundaries and of at least the primary stress. These criteria were met by (an early version of) the CELEX word-list (Kerkman, 1986), which comprised the union of the Word List of the Dutch Language and the B-list of the Uit den Boogaart (1975) corpus, totalling just under 70,000 words. The orthographic forms had been assigned a phonemic code by a computer algorithm (Kerkhoff, Wester & Boves, 1984), and corrected by hand when necessary. The phonemic code recognizes 20 vowel phonemes, and 20 consonants, as exemplified in table I.

Table I: Dutch phoneme inventory adopted in the lexical database.

EI	reis	U	put	v	veel
AU	houdt	o	rood	s	sok
UI	muis	O	rot	z	zee
E:	serre	a	maat	S	chocola
O:	zone	A	mat	Z	jaquet
U:	freule	A:	half-time	x	lachen
i	liep	@	de	g	liggen
I	pit	p	pas	m	maat
y	fuut	b	bal	N	bang
u	boek	t	tak	l	lang
e	lees	k	kas	r	rijk
E	pet	G	goal	w	wang
&	deuk	f	fok	j	jan
				h	hand

The great majority of the entries in this lexicon are morphologically complex, comprising both derivations and compounds but no inflections. Unfortunately, our version of the lexicon did not indicate morpheme boundaries. Since the inventory of prefixes and suffixes is rather small in Dutch, the high lexical frequency of segment sequences coinciding with such affixes might be at variance with the distribution of segments in word stems. It is therefore necessary to be able to isolate the monomorphemic entries in the lexicon. To this end we submitted the entire lexicon to a morphological decomposition routine (MORphological PArser, de Haan & Paerels, 1984) with a 12,000 item morpheme lexicon. The procedure was adapted such that each input word was

classified by a strictly binary decision as either monomorphemic or complex. Words that could not be parsed by the algorithm, were analysed by hand.

3. Analyses and results

Distribution of syllable types in initial and final position

Since our echoic memory contains only 1/4 second of sound, or roughly one syllable (cf. introduction), it makes sense, as a first approximation, to examine the distribution of contrasts in word-initial syllables, and compare this with word-final syllables. If it is true that the word onset is more likely to contain information as to the word's identity, we would expect that the number of different syllables that can appear in word intial position, exceeds the number of different word final syllables. Using the lexicon described above as our database, we generated a complete inventory of Dutch syllable types broken down into four categories as indicated in table IIa. Category (i) contains syllable types that occur exclusively in word initial position, category (ii) occurs exclusively in word final position, and category (iii) only in word medial positions. Category (iv) contains those syllable types whose occurrence is not restricted to a single word position.

Table IIa: Absolute and relative lexical frequencies of syllable types in Dutch, broken down by four distributional categories (see text). Prosodic differences between syllables have been ignored.

abs.	rel.	distribution
2032	(28%)	exclusively word initial
1415	(19%)	exclusively word final
687	(9%)	exclusively word medial
3207	(44%)	no specific distribution
7341	(100%)	total

Crucially, when comparing the top two rows in this table, we observe that word-initial syllable types clearly outnumber the word-final types.

So far, however, syllables have been considered different only if they differed in one or more phonemes; differences between stressed and unstressed vowels have been ignored. Let us therefore include stress information as a contrastive element differentiating among syllable types, as has been done in table IIb.

Table IIb: As table IIa, but stressed and unstressed variants of vowels are accepted as contrastive elements.

abs.	rel.	distribution
2865	(27%)	exclusively word initial
1715	(16%)	exclusively word final
757	(7%)	exclusively word medial
5342	(50%)	no specific distribution
10683	(100%)	total

Notice, first of all, that the absolute number of syllable types has increased by about 50%, indicating that roughly half of the syllable types listed in table IIa occur twice in the Dutch lexicon: once stressed and once unstressed. Stress therefore provides, at least potentially, a powerful cue to distinguish between words in the lexicon.

Secondly, we observe once more that the inventory of different word-initial syllables is richer than the word-final inventory. Most importantly, the predominance of contrasts in initial syllables is more pronounced when stress is added as a distinguishing feature. The number of initial and final syllable types in table IIa (2032 vs. 1415, respectively) is more evenly distributed than in table IIb (2865 vs. 1715), chi square = 16.6 (df = 2), p = 0.001.

The functional load of the stressed/unstressed contrast is higher in initial syllables than in final syllables. Therefore the distribution of stress position in the lexicon seems to be organised so as to help differentiate between alternative recognition candidates at the earliest possible moment.

Distribution of stress patterns in Dutch word types

Comprehensive frequency data on stress pattern distribution have never been published for Dutch. In this section we shall therefore examine the distribution of stress patterns in our version of the CELEX word-list. By stress pattern we shall mean the rhythmic shape of a word expressed in terms of its length in number of syllables and the position of the (primary) stress within the array. Table IIIa presents the distribution of stress patterns for monomorphemic words in the Dutch lexicon. Just over 12,000 entries in our 70,000 word lexicon were listed as monomorphemic.

Table IIIa: lexical frequency of stress patterns in Dutch monomorphemic words. Cell percentages are relative to row totals.

vertically:	word length in syllables								
horizontally:	stress position								
	1	2	3	4	5	6	7	8	total
1	4284								4284
	100%								
2	2703	1682							4385
	62%	38%							
3	408	808	1032						2248
	18%	36%	46%						
4	56	128	474	314					972
	6%	13%	49%	32%					
5	7	-	43	76	52				178
	3%		24%	43%	29%				
6	-	-	-	6	9	5			20
				30%	45%	25%			
>6	-	-	-	-	2	-	-		2
					100%				
	7458	2618	1549	396	63	5	-	-	12089

It appears from these data that, in monomorphemic Dutch words, stress generally falls within the final three syllables, with a modest preference for the penultimate position. This statistical distribution is quite adequately predicted by the stress rules proposed by metrical phonologists (Don & Zonneveld, 1988, and references given there; Langeweg, 1988). Though a few monosyllabic function words are unstressable (not indicated in table IIIa), they constitute less than 0.5% of the monosyllables, and hence are not reflected in the table.

Table IIIb presents the data for the complete lexicon, collapsed over monomorphemic and complex words. Table IIIb is not fully comparable with table IIIa. In the CELEX-lexicon verbs are listed as infinitives, i.e., as stems followed by an inflectional ending consisting of a single schwa. However, most stem-final consonants will be resyllabified with the inflectional ending. In our monomorphemic lexicon, verbs were listed as stems only. For instance, in the monomorphemic lexicon there is a verb breng /brEN/ that is absent in the CELEX-list, where it occurs only in in vin-den /vIn-d@/. As a result, there are more monosyllables in table IIIa than in table IIIb. After this caveat, let us consider the figures.

Table IIIb: As table IIIa, but data accumulated over the entire lexicon

vertically: word length in syllables
horizontally: stress position

	1	2	3	4	5	6	7	8	total
1	3373								3373
	100%								
2	15758	2726							18484
	85%	15%							
3	18020	6370	2606						26996
	67%	24%	9%						
4	6436	3365	3036	1278					14115
	45%	24%	21%	10%					
5	1532	1016	928	927	361				4764
	32%	21%	19%	19%	8%				
6	347	288	279	112	173	77			1276
	27%	23%	22%	9%	14%	6%			
7	75	53	64	48	21	21	13		295
	25%	18%	22%	16%	7%	7%	5%		
8	13	9	17	12	8	-	5	3	67
	19%	13%	25%	18%	12%	-	7%	5%	
9	2	1	1	1	4	1	-	-	10
	20%	10%	10%	10%	40%	10%			
>9	-	-	-	-	2	-	-	-	2
					100%				
	45556	13828	6931	2378	569	99	18	3	69382

As may be observed in table IIIb, the primary stress occurs in virtually any position within the word when complex words are included in the lexicon. Since stress is most likely to fall on the initial part of a Dutch compound word (Langeweg, 1988), which is the most frequent type of complex word in our lexicon, the clear preference for stress on an early syllable is predictable.

This statistical distribution of stress positions over word length may assist in efficient and successful word recognition in at least the following two ways:

(i) When the target word is still being spoken, the stress information may guide the listener's decisions in eliminating unlikely recognition candidates and (de-)activating specific sublexicons. For both monomorphemic and complex words, roughly two out of every three begin with a stressed syllable. Therefore, especially hearing an unstressed word onset should allow the listener to exclude a large portion of the mental lexicon from the revelant search space.

(ii) When the entire rhythmic pattern is available to the listener, i.e. after the spoken word has been completed, the lexical search space is severely limited. If the listener has not yet recognized the word at this point, for instance when the speech is acoustically impoverished, the largest sublexicon that has to be searched comprises trisyllabic words with initial stress. This sublexicon is less than a quarter of the entire lexicon. For all other rhythmic patterns the lexical search space is even smaller.

Distribution of lexical recognition points

According to the so called cohort model of spoken word recognition, words will be recognized at the earliest possible moment (Marslen-Wilson, 1985). When a word is presented out of context, recognition will take place at the lexical uniqueness point (UP), the place within the word where it is first uniquely distinguished from all other words in the lexicon. For instance, the UP for the word elephant is reached at the fourth phoneme, [f], where it is first distinguished from e.g. element: there are no other words in English that begin with the sound sequence [eləf...] than precisely elephant (and its derivations).

If it is true that the organisation of the lexicon is such that words are distinguished more efficiently in their beginning sounds, one would predict that the UP is reached sooner when going from left-to-right than from right-to-left. Using the same example, the UP for elephant analysing the lexicon from right-to-left (backwards) is reached at [...əfənt] where [ə] distinguishes it from e.g. infant: there is no English word other than elephant that ends in [...əfənt]. In this example the forward UP lies 4 phonemes from the word onset, but the backward UP at 5 phonemes from the word ending. Table IV contains the results for Dutch as we computed them for our version of the CELEX word-list.

We conclude from this table that, on average, the UP is not reached sooner from the left than from the right on a purely segmental basis. When stress information is allowed to contribute to the word's identity, we notice, first of all, that the UP is reached about 1 phonemic segment earlier. Crucially, the acceleration due to stress information is larger when words are analysed from left-to-right than vice versa. These effects are qualitatively the same as those reported for other Germanic languages, in particular for Swedish, English, and German (Carlson et al., 1985).

Although this asymmetry supports our position, the effect is disappointingly small. Therefore we propose yet another, hopefully more revealing, analysis of the distribution of contrasts in the lexicon.

Table IV: Mean position of lexical Uniqueness Point measured from left-to-right (from word onset) and from right-to-left (from word ending) with and without inclusion of stress as a distinctive characteristic. The data have been accumulated over the entire lexicon including monomorphemes and complex words.

	Without stress information	With stress information
Mean word length in phonemes:	8,6	8,6
Mean Uniqueness Point (from word onset)	6.9	5.7
Mean Uniqueness Point (from word ending)	6.8	6.0

Reduction of cohort size

Going through the word forwards or backwards does not affect the average position of the lexical UP. For all this, we did observe that initial syllables are more diversified than final syllables. Therefore it seems reasonable to expect that the number of recognition candidates (the cohort size) shrinks faster when going from left-to-right than vice versa, so that at any comparable position in the word, there are fewer possibilities for the listener to choose from when going from left-to-right. As a general rule, word recognition will be easier as there are fewer alternatives to choose from.

The relevant descriptive statistic is rather complicated. It should not be difficult to appreciate that simple measures, such as mean cohort size as a function of fragment length, are inadequate. For instance, on the basis of an onset fragment of just 2 phonemes, as many as 484 different cohorts are obtained, each containing 143 words on average, but ranging in size between 1 and 2144 words. We argue that the listener's uncertainty as to the intended word is most adequately expressed by a measure called Entropy (H) in information theory (cf. van Heuven, 1978 and references given there; Shannon, 1949), defined as:

$$H = - \Sigma\ p_i\ {}^{2}\log\ p_i,$$

where i is an index ranging over all the cohorts under consideration, e.g., 484 in the above example, and where p_i is the proportion of a cohort relative to the entire lexicon. When the length of the word fragment is 0 (i.e., no phoneme has been given yet), $H = {}^{2}\log\ 69{,}382 = 16.08$ bit. When the word fragment approaches the length of the longest word in the lexicon, H will rapidly decrease to 0. Roughly, entropy expresses the average number of binary divisions of the search space (in bits) necessary to locate a single element. Reduction of entropy by 1 bit reduces the number of alternatives to choose from to 50 per cent. The results are as in table V.

It is quite clear from the entropy data that the cohort size is reduced much more efficiently going forward from the word onset than going backward from the word ending. During the first 4 phonemes (i.e., roughly one syllable) the listener's uncertainty as to the word's identity is 1 bit less going forward than going backward; or, stated differently, the number of alternatives to

choose from when going from left-to-right is systematically smaller (by 50%) than when going from right-to-left. After 4 phonemes from the leading word edge the listener has $2^{3.52}$ = just over 11 words, on average, to choose from. In combination with syntactic and semantic information derived from the preceding context, the word will practically always be available at this point.

Table V: Entropy (in bits) as a function of sound position, from word onset versus word ending.

Sound position	From word onset	From word ending	Difference
0	16.08	16.08	--
1	11.68	12.56	0.88
2	8.49	9.53	1.04
3	5.69	6.77	1.08
4	3.52	4.48	0.96
5	2.03	2.64	0.61
6	1.07	1.43	0.36
7	0.56	0.74	0.18
8	0.30	0.38	0.08
9	0.15	0.18	0.03
10	0.08	0.08	0.00
11	0.04	0.04	0.00
12	0.02	0.02	0.00
13	0.01	0.01	0.00
14	0.00	0.00	0.00

4. Conclusions and discussion

Taking our cue from insights into the process of spoken word recognition, we have examined aspects of the structure of the Dutch lexicon. If language is optimally adapted to the perception of speech, rather than print, we expect contrastive elements to cluster in the early parts of words. Secondly, it was an open question to what extent prosodic information, notably stress, might assist in establishing word identity from shorter (initial and final) word fragments. Finally, we asked whether the distribution of segmental and prosodic contrasts would be different for morphologically simple versus complex words.

Our results indicate (table II) that the Dutch lexicon indeed concentrates segmental contrasts towards the word onset. The number of different syllables that occur at the beginning words is clearly larger than at the end of words. Moreover, the advantage of the onset syllable increases considerably if stress is included as a discriminating feature.

On average (Table IV), a word can be identified in our lexicon after that 80% of the phonemes have been used, counting from the leading word-edge, or 79% from the trailing edge. When stress information is included, a forward search is already successful after 66%, on average, whereas a backward search is successful after 70%. Inclusion of stress therefore allows the identification of words in the lexicon from shorter fragments. Curiously enough, however, the position of the lexical uniqueness point is hardly affected by the direction of the search.

Cohort size shrinks faster during forward search than during backward search (table V). During the first 4 phonemes the lexical search space is consistently 50% smaller during forward search than during backward search. The striking advantage of the forward search disappears rapidly after the fourth segment, and is practically 0 by the time the lexical uniqueness point has been reached.

Finally, there were no indications that the phonemic structure of morphologically complex words differs from that monomorphematic words.

There is a lot of evidence in the literature to suggest that spoken words are recognized more effectively from onset fragments than from equally long final portions (e.g., Nooteboom, 1981; Salasoo & Pisoni, 1985). This finding seemed to be in line with the special status accorded to the word onset in recognition models described by Cole & Jakimik (1978, 1979) and Marslen-Wilson (1985).

The results of our survey of statistical properties of the Dutch lexicon, and of related languages by Carlson et al. (1985), indicate that these experimental data do not necessarily require the postulation of a processing mechanism that directs special attention to the beginning of words. The superiority of the word onset in recognition experiments can now be explained in an alternative fashion: the superiority of the word onset is simply due to its greater functional load. Crucially, in a series of experiments where the lexical material was carefully selected so as to control for the asymmetry in lexical density between word beginning and ending, no traces of the word onset superiority remained (van der Vlugt, 1987).

8. References

BOOGAART, P.C. UIT DEN (ed)
1975 Woordfrequenties in geschreven en gesproken Nederlands, Utrecht, Oosthoek, Scheltema en Holkema.

CARLSON R., ELENIUS, K., GRANSTRØM, B., HUNNICUT, S.
1985 Phonetic and orthographic properties of the basic vocabulary of five European languages, in Speech Transmission Laboratory - Quarterly Progress and Status Report, 1, 63-94.

COLE, R.A., JAKIMIK, J.
1978 Understanding speech: how words are heard, in G. Underwood (ed) Strategies of information processing, New York, Academic Press.

COLE, R.A., JAKIMIK, J.
1979 A model of speech perception, in R. Cole (ed) Perception and production of fluent speech, Hillsdale NJ, Erlbaum.

CUTLER, A.
1987 Forbear is a homophone: lexical prosody does not constrain lexical access, in Language and Speech, 29, p. 201-220.

DON, J., ZONNEVELD, W.
1988 VC-phonology, theory and machine in Dutch stress assignment, in Progress Report of the Institute of Phonetics Utrecht, 13.1, p. 8-32.

HAAN, M. DE, PAERELS, M.
1984 Morpa, een morfologische ontleder [Morpa, a morphological parser], unpublished report, Dept. of Computer Science/Phonetics Laboratory, Leyden University.

HEUVEN, V.J. VAN
1978 Spelling en lezen [Spelling and reading], Assen, van Gorcum.

KERKHOFF, J., WESTER. J., BOVES, L.
1984 A compiler for implementing the linguistic phase of a text-to-speech conversion system, in H. Bennis, W.U.S. van Lessen Kloeke (eds) Linguistics in the Netherlands 1984, Dordecht, Foris, p. 111-117.

KERKMAN, H.
1986 Voorlopige beschrijving Celex-bestand [Provisional description of the Celex database], unpublished report, Interfacultary Working Group Language and Speech Behaviour, Catholic University Nijmegen.

LANGEWEG, S.J.
1988 The stress system of Dutch, doctoral dissertation, Leyden University.

MARSLEN-WILSON, W.D.
1985 Spoken word recognition: a tutorial review, in H. Bouma, D. Bouwhuis (eds) Attention and Performance, X, London, Erlbaum, p. 125-150.

NOOTEBOOM, S.G.
1981 Lexical retrieval from fragments of spoken words: beginnings versus endings, in Journal of Phonetics, 9, p. 407-424.

SALASOO, A., PISONI, D.
1985 Interaction of knowledge sources in spoken word identification, in Journal of Memory and Cognition, 2, p. 210-231.

SHANNON, C.E.
1949 The mathematical theory of communication, in C.E. Shannon, W. Weaver (eds) The mathematical theory of communication, Urbana, The University of Illinois Press, p. 3-91.

VLUGT, M. VAN DER
1987 Spraakgeluid en woordherkenning: het relatieve gewicht van het begin en eind van een gesproken woord [Speech sound and word recognition: the relative weight of the beginning and ending of a spoken word], doctoral dissertation, Technical University Eindhoven.