EFFECTS OF STRESS AND ACCENT ON THE HUMAN RECOGNITION OF WORD FRAGMENTS IN SPOKEN CONTEXT: GATING AND SHADOWING

Vincent J. van Heuven

Phonetics Laboratory/Dept. of Linguistics
Leyden University, P.O. Box 9515
2300 RA Leiden, The Netherlands

1. INTRODUCTION

One of the current issues in auditory word recognition concerns the role of lexical stress. In languages such as English and Dutch, the position of the stressed syllable varies from one word to the next, and may therefore supply information that helps the listener to recognise the word. However, the experimental data so far present a confusing picture.

Researchers have been quite successful in showing that word recognition suffers if stress is pronounced on the wrong syllable, be it unwittingly by foreigners [1] or deliberately by native speakers [2,3,4], or if word stress is not marked at all, as is possible in synthetic speech [5]. Also, poorly defined segmental information is reinterpreted by the listener so as to fit the perceived stress pattern [1,6].

The recognition of correctly pronounced words does not improve in any way when the listener is given advance information on the stress pattern, even though this information may help the listener to pre-activate a specific portion of the lexicon [2]. Cutler [7] goes further still, arguing that stress information does not play any role at all in lexical access, i.e., during the early stages of word recognition. In a cross-modal priming task subjects made lexical decisions for a visually presented target which was or was not synonymous with a prime word embedded in a concurrent, spoken utterance, preceding the visual target by either a short (250 ms) or a long (750 ms) lag. Primes were minimal stress pairs of the type FORbear ('ancestor') versus forBEAR ('endure'). For short prime-target intervals lexical decision was speeded up for both readings of the prime words, from which Cutler concludes that minimal stress pairs are homophonous during lexical access. Only for long prime-target intervals did one reading of the prime (the semantically related one) but not the other, facilitate the recognition of the target, showing that the effect of stress must be post-lexical. In this view lexical access is based on segmental information alone.

It seems odd, however, that listeners should ignore an information source that is obviously present in the auditory signal, that might help them to reduce the number of recognition candidates during lexical access. There is a wealth of experimental data suggesting that listeners use any bit of bottom-up information at the earliest possible moment (cf. [8]). Therefore the purpose of the present experiments was to gather more evidence on the possible role of lexical stress during the early stages of word recognition, using different experimental techniques than Cutler's cross-modal priming.

We reasoned that a contribution of lexical stress could be demonstrated as follows. As long as a word cannot be uniquely distinguished from its competitors in the lexicon, lexical access is still in progress. Suppose we cut off an utterance at such a point in a polysyllabic target word that the information of the context and the initial segments of the target leave exactly two recognition candidates, one with and one without a stressed onset syllable. If it is true that stress in not used during lexical access, listeners will come up with the same distribution of responses, whether the target onset is stressed or not. If, on the other hand, stress does play a role in access, listeners should be able to isolate precisely one word, viz. the alternative matching the stress characteristic of the stimulus fragment.

Cutler [7] points out that the contribution of stress to lexical access can only be studied properly in cases where stress does not correlate with a segmental difference (e.g., full vowel quality versus reduction to shwa). In order to make sure that the effects we examine in the present experiment are really prosodic (rather than segmental), we decided to systematically vary the strength of the prosodic stress cues by having the targets pronounced with and without accent. In Dutch (and in English) stressed syllables in an accented word are pronounced with a conspicuous pitch movement, which is by far the most salient cue for stress (cf. [9]). The stressed syllable in a de-accented word loses its pitch movement (e.g. [10]), so that stress position is cued much less clearly. Therefore, if, in the experiment we propose to undertake, the choice between the two alternatives is made more successfully for accented targets, the effect of stress on lexical access must be prosodic in nature rather than segmental.

2. EXPERIMENT I

2.1 Method

As a first approximation we adopted the so called gating paradigm [11] as our experimental method. This technique simulates lexical access in an off-line fashion by presenting a word not just once, but repeatedly. On the first pass only a short (initial) portion of the spoken word is made audible, and the listener is asked to guess what word he has heard the beginning of. On each successive presentation a larger portion of the word is made audible, and each time the listener is asked to revise his guess. In the responses we can trace (i) the length of initial stimulus portion that is necessary for successful completion of the target, and (ii) the narrowing-down of the pool of recognition candidates as the audible fragments grows longer.

We selected 8 Dutch word pairs, all of which were polysyllabic, monomorphemic nouns. The members within each pair had at least the same onset CV-combination, and were matched for number of syllables. Crucially, one member of the pair had a stressed first syllable, whereas the other bore the lexical stress on the second syllable, leaving the first syllable unstressed.

For each target pair a context sentence was constructed that, in combination with the segmental information contained in the onset CV, narrowed down the number of alternatives to exactly the two members in the pair. The following is an example:

In een kooitje in de hoek van de kamer zat een {KAvia / kaNArie}
[in əŋ ko:jcə in də hu:k fan də ka:mər zat əŋ {'ka:vi:a:/ka:'na:ri:}]
'in a cage in the corner of the room sat a {cavy / budgy}'

Each context-plus-target combination was recorded twice by the author. In one version the target, which was the sentence-final noun, was invariably accented. In the second reading the target was de-accented, while the speaker realised a contrastive accent on an earlier word in the sentence.

The 32 recordings were A/D converted (10 kHz, 12 bits, 4.5 kHz LP) and stored on disk. Using a digital waveform editor, the first truncation was made at the latest point in the first syllable of the target word where the members of a pair did not differ segmentally. For instance, in the word pair ORgel (organ) versus orKEST (orchestra) the cut was made in the [r] just before the onset of the closing gesture towards [ɣ, k]. The second gate included the complete portion of the waveform corresponding to the earliest segment that distinguished the competing words; in the orgel/orkest-case this included either the full fricative [orɣ] or the the full velar plosive plus release [ork^h]. The third gate comprised the context sentence plus the whole target

The fragments were D/A converted and recorded onto 4 different tapes. Each tape contained the 8 different context sentences in the same random order with the 3 gates of increasing length separated by ISI's of 5 seconds (offset to onset). Each tape contained only one member of each target pair, with stress position and accent conditions blocked over tapes and subjects.

The 4 tapes were played to small groups of subjects (10 per tape). Subjects were instructed to guess the identity of the last word of each sentence they heard on the tape, and to select polysyllabic words only.

2.2 Results

Each word response was scored for segmental and rhythmic correctness. A response was segmentally correct if its initial phonemes perfectly matched the segmental make-up of the audible portion of the target. Responses were scored rhythmically correct if the stressed/unstressed nature of the initial target syllable was identical to that of the stimulus. In table I the data are first broken down into legal and illegal responses. Legal are only those responses that are identical to the competing target pairs. Legal responses are further broken down into rhythmically correct and incorrect responses. All other responses are illegal. The data for accented and de-accented targets, collected for gate #1 only, are presented in separate columns.

Table I: Number of responses at gate #1.

RESPONSE TYPES	ACCENTED TARGETS			DE-ACCENTED TARGETS		
	stress on syllable		total	stress on syllable		total
	#1	#2		#1	#2	
1. segments correct, stress correct	37	38	75	21	45	66
2. segments correct, stress wrong	14	10	24	25	7	32
subtotal: legal responses	51	48	99	46	52	98
3. illegal responses	29	32	61	34	28	62
total responses	80	80	160	80	80	160

Let us first discuss the data obtained for accented targets. Our original expectation had been that subjects should be able to narrow down the response set to just the two alternatives that both fitted in the semantic context and matched the segments contained by the first gate. From this perspective illegal responses should not have occurred at all. Yet it is apparent from table I that such responses did occur rather often, indicating that the task was far from easy: there are 61 illegal responses (38%), 30 of them being segmentally correct monosyllables (19%). Segmental intelligibility was adequate, with 81% correctly reported segments even at the first gate (results not indicated in table I). Ninety-nine of the 160 possible responses (62%) were legal, i.e., properly constrained to the competing minimal stress pairs. Crucially, the choice between the alternatives within the pairs was correctly constrained by the stress cues in the audible fragment in the great majority of the cases: 75 out of 99 cases (76%), with essentially the same distribution for targets with and without initial stress: 37/51 (or 73%) versus 38/48 (or 79%), respectively, $X^2(1)=.59$ ($p > .30$). In either stress condition the distribution of correct responses is highly significantly above chance with $z=4.27$ ($p<<.001$; one-tailed binomial test) for stressed onsets, and $z=4.04$ ($p<<.001$) for unstressed onsets.

The alternatives reach their segmental recognition point (as defined in context) at gate #2. Here we expect (near-)perfect completion of the target fragments, irrespective of stress information. This is indeed the case, with 114 correct responses out of 120 legal responses (= 95%). However, the task is still difficult, given that 40 out of 160 possible responses are illegal

(25%). Recognition does not reach perfection until gate #3, with 99% correct responses.

Let us now examine the data collected for unaccented targets. Remember that lexical stress is cued here by temporal organisation (lengthening for stress, shortening for no-stress) only, since the speaker omits pitch movements in order to signal de-accentuation. Consequently we expect weaker differentiation between stressed and non-stressed word onsets.

We observe, first of all, that there is virtually no difference in segmental differentiation. Again 81% of the responses were segmentally correct (not in table), and 61% were legal. Within the category of legal responses, the effect of lexical stress, though still significant, is much weaker here than in the accented targets: the correct alternative is selected in 66 out of 99 legal cases (67% correct with 50% chance). When the onset syllable is unstressed, the correct alternative is selected in 45 out of 52 cases (87% correct), which is far better than could be obtained by chance, $z=5.13$ ($p<<.001$). However, when lexical stress is in initial position in de-accented targets, there is still a slight majority of responses that select the alternative with non-initial stress (25 out of 46 cases = 54%). Although this distribution of responses is essentially random, $z=.44$ ($p > .2$), the proportion of initially stressed responses is much larger here than for unstressed onsets, $X^2(1)=12.4$ ($p<.001$).

After the second gate the proportion of legal responses is 74%, which is about equal to that obtained for accented targets (75%), but the constraining power of the prosody has dropped slightly: 102/119 (86%) correct here versus 114/120 (95%) for accented targets, $X^2(1)=5.9$ ($p<.05$). At gate #3 all the responses, except for 4 blanks, are correct (98%).

2.3 Conclusions and discussion

The results of this experiment demonstrate that stress information does play a role in lexical access, i.e., during the early stages of the word recognition process when the recognition system is still gathering information to narrow down the set of recognition candidates. Our listeners proved able to integrate top-down information from the preceding context and bottom-up information provided by the audible portion of the sentence-final target word, even when the latter was truncated halfway through its onset syllable. The stimuli were constructed such that the segmental information at gate #1 could not narrow down the set of alternatives to just one; the choice could be uniquely constrained only if the listener took prosody into account.

Obviously then, stress is used effectively in lexical access, helping the listener to select the one and only correct alternative at the earliest possible moment, i.e., even before the remaining alternatives are segmentally distinct. Therefore, our results do not support Cutler's [7] claim that stress does not play a role in lexical access. It is difficult to decide whether our results falsify her claim. This decision depends on the valitidy of the gating method as a simulation of the on-line recognition process. It is generally accepted that the results obtained in a gating experiment set the upper boundary to what listeners can do with bottom-up information. If certain cues are effective in a gating task, this may be due to the fact that the listener is allowed more time to consider his response. Therefore, a follow-up experiment will presented in our next section to determine how much of the potential information is actually used in on-line word recognition.

3. EXPERIMENT II: SHADOWING

3.1 Method

In our second experiment we have tried to replicate experiment I using an on-line word recognition task. It is widely accepted that fluent shadowing, i.e. immediate verbal repetition by the listener of a spoken utterance or text, is a clear case of on-line word recognition (e.g. [8]).

As before, we presented stimuli that contained a stressed or unstressed first syllable embedded in a context utterance that constrained the target to two alternatives, which could be distinguished on rhythmic grounds only. In order to enable our subjects to continue shadowing the stimulus fluently, targets were embedded in the context utterance in medial rather than final position. The second part of each target, containing segmental differences within the pairs (cf. experiment I), was replaced by pink noise, so that all segmental differences between the alternatives were eliminated from the stimuli. The noise was loud enough to cause the subjects to believe that the utterance was never interrupted (cf. [12]).

If our hypothesis is correct that stress information is used even in the early stages of word recognition, the shadower should choose the rhythmically correct alternative significantly more often than can be expected on a chance basis.

The 2 * 8 sentences were recorded on audio tape by a female speaker of Dutch, using, and then stored in computer memory. The final portion of each target initial syllable, as well as the remainder of each target word, was gated out (using the same criteria and procedure as before), and replaced by a steady state pink noise burst of constant duration. Stimuli were then re-recorded on two audio tapes, where each tape contained 4 stressed and 4 non-stressed target onsets, with the alternatives per pair blocked over tapes.

Two groups of 11 male Dutch speakers were instructed, after suitable practice, to repeat the messages on the tape as they were listening to it, promptly and fluently, avoiding omissions and hesitations, and ignoring noise bursts as well as they could.

3.2 Results

The subjects' responses were first analysed for disfluencies. Two skilled judges decided independently of one another whether a response contained an omission, hesitation, decrease in speaking rate, or any other audible form of disfluency. Only when both judgments agreed were responses included in our further analysis. Table II shows the results in absolute and relative frequencies, broken down into the relevant categories (cf. experiment I).

Table II:
Number of responses to mutilated accented targets in a shadowing task.

RESPONSE TYPES	stress on syllable		total
	#1	#2	
1. segments correct, stress correct	58	33	91
2. segments correct, stress wrong	5	9	14
subtotal: legal responses	63	42	105
3. illegal responses	33	54	87
total responses	96	96	192

As expected there are many disfluencies; yet in 55% (105/192) of the cases fluent shadowing was obtained where the choice was adequately constrained to the members of the crucial pairs. It is quite apparent from the data that the stress information in the target onset provides useful information to the listener. The correct alternative is chosen in the great majority of the legal cases: 92% for stressed onsets and 79% for unstressed onsets, ($p<<.001$, one-tailed binomial test). There is no significant difference between the proportion of correct responses obtained for stressed and unstressed onsets, $X^2(1)=2.9$ (ins).

3.3 Conclusions

In more that half of the responses to our stimuli we obtained fluent shadowing, where word recognition is on-line, with the choice between candidates adequately constrained to the members of the quasi-minimal stress pairs. For this group of data stress information provides the vital cue to the subjects to select the one and only correct alternative, which was then chosen in no less than 87% on average. This proportion of correct responses is in fact larger than that in experiment I, where the contribution of stress information to word recognition was examined in an off-line task.

The results of experiment II therefore strongly support our claim that prosodic information, notably the difference between stressed and unstressed but segmentally identical word onsets, is used in the word recognition process, not only in off-line tasks, but also in the lexical access phase during on-line recognition.

Summing up then, our results do show that listeners can use stress information during lexical access, when they are forced to do so. In our stimuli stress provided non-redundant information (at gate #1), that was vital to the early isolation of the target. In Cutler's experiment stress information was redundant: the prime words could be accessed from context and segmental information without having recourse to prosodic cues. Possibly, since her subjects had no need to take prosody into account, both members of the minimal stress pairs were activated in the mental lexicon. Had Cutler used non-constraining contexts, so that stress position had to be used in order to resolve an ambiguity, she might well have found that minimal stress pairs are not homophones after all.

4. REFERENCES

[1] R K BANSAL, The intelligibility of Indian English, Ph.D. thesis, London University (1966)

[2] A CUTLER & C E CLIFTON, 'The use of prosodic information in word recognition', in H. Bouma, D.G. Bouwhuis (eds.): Attention and Performance X, Erlbaum, Hillsdale, NJ., p183 (1984)

[3] V J van HEUVEN, 'Perception of stress pattern and word recognition: recognition of Dutch words with incorrect stress position', JASA, 78, S21 (1985)

[4] L M SLOWIACZEK, 'Effects of lexical stress placement on auditory word recognition', paper presented at the 112th Meeting of the ASA, Anaheim, CA (1986)

[5] S G NOOTEBOOM & G J N DOODEMAN, 'Cues for lexical stress recognition of polysyllabic words, synthesised from diphones and presented in isolation', paper presented at the 109th Meeting of the ASA, Austin, TX (1985)

[6] C M CONNINE C CLIFTON & A CUTLER, 'Effects of lexical stress on phonetic categorization', Phon., 44, p133 (1987)

[7] A CUTLER, 'Forbear is a homophone: Lexical prosody does not constrain lexical access', Lang. Sp., 29, p201 (1987)

[8] W D MARSLEN-WILSON, 'Speech understanding as a psychological process', in J.C. Simon (ed.): Spoken language generation and understanding, Reidel, Dordrecht, p39 (1980)

[9] A F van KATWIJK, Accentuation in Dutch, van Gorcum, Assen (1974)

[10] V J van HEUVEN, 'Stress patterns in Dutch (compound) adjectives: acoustic measurements and perception data', Phon., 44, p1 (1987)

[11] F GROSJEAN, 'Spoken word recognition processes and the gating paradigm', Perc. Psych., 28, p267 (1980)

[12] R M WARREN & C J OBUSEK, 'Speech perception and phonemic restorations', Perc. Psych., 9, p358 (1971)