Proceedings of The Institute of Acoustics

EVIDENCE FOR RIGHT-TO-LEFT FLOW OF INFORMATION IN HUMAN SPEECH PERCEPTION

E.G. Bard (1) (2), R.C. Shillcock (2), Gerry T.M. Altmann (2)

(1) Centre for Cognitive Science, University of Edinburgh
(2) Centre for Speech Technology Research, University of Edinburgh

INTRODUCTION

This paper addresses the assumption that words in running speech are recognized one-by-one and left-to-right, with the interactive aid of prior, or "left", context only. We will claim that context following a word's offset ("right context") is frequently implicated in the interactions which select among acoustically-based word candidates.

Left-to-right models of lexical access [1-4] propose that word recognition can be achieved on-line and on-time because the interaction of higher level knowledge with the "cohort" of word hypotheses is so efficient that a word may be recognized *before* its acoustic representation is completely available to the listener. Early recognition allows such models to circumvent the difficulties of segmenting the speech stream into words on acoustic grounds alone: if word recognition coincides with or precedes the end of a word, the processor can identify the next word onset as the beginning of the acoustic material not accounted for within the word just recognised [4].

Evidence exists, however, to suggest that this type of model is an idealisation of real speech processing. Several studies have demonstrated that words are not always recognised within their acoustic lifetime [5-8], and consequently that rightwards information may be recruited during the disambiguation of the acoustic wordshape. These studies consisted of small-scale experiments which could not demonstrate, however, the generality of the phenomenon as it occurs in the perception of continuous speech.

Evidence for the absence of right-to-left information flow comes from experiments using well articulated, read speech. Such materials represent only one end of a naturally occurring continuum of articulatory clarity, while most speech lies somewhere towards the other, less well-articulated, end of this continuum. Carefully articulated experimental materials are therefore likely to provide underestimates of the role of contexts. Accordingly the present experiments employ spontaneous conversational speech to assess the generality of the right context phenomenon.

If the right-context effect is a general phenomenon, this has considerable implications for a theory of word recognition. If every word presented is recognized one trial late, that is, on its second presentation, recognition may nonetheless proceed without the use of any subsequent context which the listener has identified. Recognition may be proceeding word-by-word, though slowly, or it may be proceeding in larger units, but still with the aid only of prior context. If, on the other hand, late recognitions break presentation order and follow the recognition of

some temporally subsequent word, then the model must allow the possibility of true right-to-left information flow. Grosjean [7] has reported instances of both phenomena for monosyllablic nouns read in neutral left contexts.

In the experiments that follow, we aim to determine 1) to what extent words can be recognised given only their prior sentence contexts and acoustic shapes, 2) whether the patterns of late recognition demand right-to-left mechanisms, and 3) whether the addition of large amounts of prior conversational context can reduce listeners' difficulties and reestablish the adequacy of exclusively left-to-right information flow.

EXPERIMENT 1

Purpose

This experiment tests the hypothesis that words are recognised with reference only to their prior sentential context. It uses Pollack and Pickett's [5-6] word-level gating technique to generate normative data on a large pseudo-random sample of conversational speech. Recorded utterances are presented in gates, or continuous substrings, all starting with the first word of the sentence and increasing in length by one additional word on each successive stimulus. Listeners' failures to identify a new word on its first presentation with only its prior context are taken as counterevidence to the claim that acoustic wordshape and left context are sufficient for word recognition.

Method

Materials. Twelve utterances were randomly selected from each of 12 male and 12 female speakers for whom conversations had previously been recorded and transcribed [9-10]. The 288 utterances ranged in length from 1 to 22 words, with a mean of 6.7. Filled pauses, false starts and contracted forms (*I'll, didn't*) were all counted as single words. Stimuli were digitised at 10,000 Hz through a low-pass (5,000 Hz) filter. Word boundaries were determined so as to include as much of the preceding word as possible without including any material which could be identified as belonging to the following word. Each of four tapes represented each of the twenty-four speakers with a different set of three randomly chosen utterances, blocked and presented after an ungated orienting utterance.

Subjects and Procedure. Subjects were 48 students at University of Edinburgh with at least two years residence in Edinburgh. Twelve subjects heard each tape. They were instructed to write down the words contained in each gate, changing answers on new trials if necessary, but never altering earlier answers.

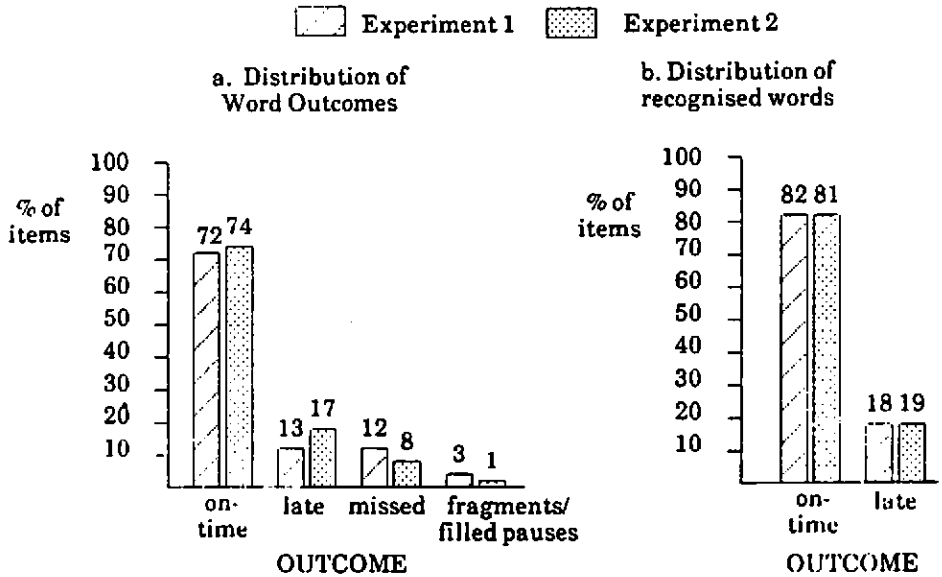
Results

The 288 utterances contained a total of 1930 word tokens, each presented one or more times to each of twelve subjects. This gives 23,160 case histories through which we can trace the time course of a listener's recognition of a word. The analysis will be framed almost exclusively in terms of these case histories.

Distribution of outcomes. Figures 1a and 1b display the distribution of outcomes by recognition point. Some 72% of all words were recognized on their first presentation with left context alone; 13% were recognized on some later presentation, in the presence of subsequent context; 12% were not successfully identified. Roughly one in every seven words identified was recognized late.

Further analysis shows that late recognitions were not an artefact of the interruption of word-final coarticulation. Approximately 54% of late recognitions were identified at least two presentations late, that is after at least one more word. The average number of additional words needed for late identifications was between one and two: thus words were presented 2.69 (s.d.=1.32) times on average.

Figure 1. Experiments 1 and 2: Distribution of recognitions by outcome.

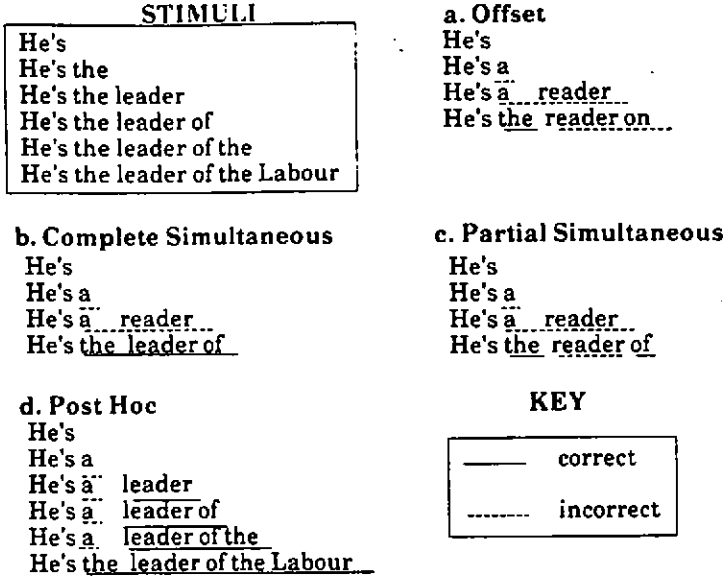


Types of late recognition. Late recognitions fall into several categories depending on the available right-context at the point of recognition. (See Figures 2 and 3.)

Category 1: Offset Recognitions. These recognitions consist of a single word recognised only when it has been presented with a subsequent context, but *before any of those subsequent words have been correctly identified.* (See Figure 2a.) This pattern provides no direct evidence that any right context assists the recognition of *the.* This type of response may receive a considerable contribution simply from the repetition of the word. Note that any number of unrecognised words may follow the Offset word when it is finally recognised. One-word Offsets may reflect the effects of coarticulation, but longer offsets are less easily explained in this way.

Category 2: Complete Simultaneous Strings. In this category a word is first recognized *on a late presentation on which all the subsequent words are also first recognized. Although at the point of recognition, the subject has* access to word identities following the word in question, one might account for the simultaneous recognition of a string simply by allowing left context to operate on the recognition

Figure 2. Types of late recognition.



of a string of words taken as a single unit. In the example in Figure 2b, *the* is recognized on the same presentation as two subsequent words and the string therefore contains three members. Simultaneous Strings which are only two words long may reflect coarticulation effects, which might be viewed as right context effects on a small scale. In the case of longer Simultaneous Strings, little coarticulation between non-adjacent words is likely and the arguments for supra-word recognition units or for a higher-level right-context effect become correspondingly stronger.

Category 3: Partial Simultaneous Strings. This category resembles the Offset category in that no right context words are identified before the word in question, but it differs in that *some word or words are recognised at the same time.* Thus there is simultaneous recognition of two or more words which do not form a complete string from the last recognised word to the end of the current stimulus. (See Figure 2c.) This category is more difficult to attribute to left-to-right processes. Some words in these sequences can be recognised while others cannot. On the other hand, correct recognition to the right of the word of interest suggests that information exists which may be flowing leftwards.

Category 4: Post Hoc Recognitions. This last category represents the strongest case for right-context effects: here *the order of word recognition is non-consecutive (see*

Figure 3. Experiments 1 and 2. Distribution of Late Recognitions.

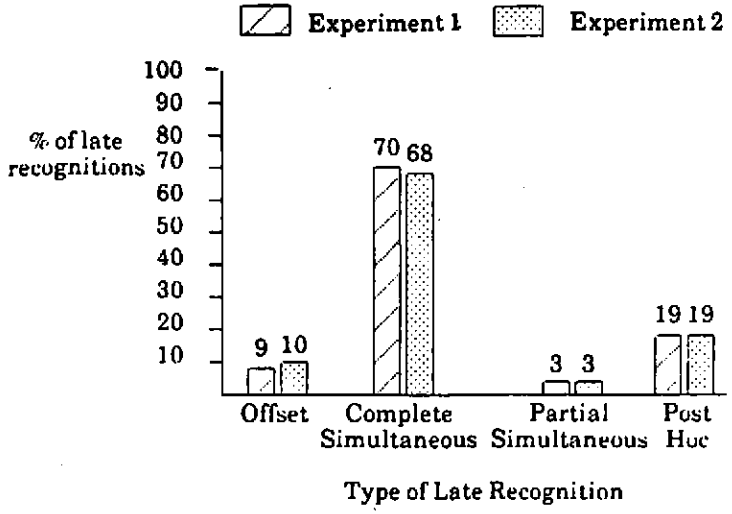


Figure 2d). suggesting that the listener has lexical or higher-level information explicitly available for disambiguating an earlier word form.

Figure 3 displays the overall distribution of late recognitions into these four categories. Separate analyses by syllable length and form-class (functor v. contentive) yield substantially the same pattern. It would seem that the mechanism mediating late recognition is not restricted to any particular subset of the mental lexicon (compare [11]).

Correlates of Recognition Outcomes. Further analyses were carried out to determine the variables contributing to these phenomena. Multiple regressions were used to examine the roles of those variables associated with the stimulus words as lexical *types* and those associated with the words as particular *tokens*.

Lexical type variables (see Table 1) were *syllable length* as determined from dictionary forms, the *functor/contentive* distinction which separates words with primarily syntactic roles from those with less restricted function, and *frequency of occurrence* [12]. Lexical token variables measured *word length in milliseconds* and *distance in words from beginning* and *end of the utterance*.

The results suggest that a word is more likely to be recognised on its first presentation if it is longer, a content word, and further from the beginning of its utterance. It is more likely to be recognised late or not at all if it is a functor,

Table 1. Experiment 1. Multiple Regressions for On-time, Late and Missed Recognition

		Dependent variable		
Independent variable		% on time	% late	% missed
TYPE	Functor/Contentive	-0.097**	0.088**	0.068*
	Syllable Length	-0.026	0.015	0.024
	Word Frequency	0.022	0.014	-0.051
TOKEN	Distance from utterance onset	0.230***	-0.117***	-0.225***
	Distance from end of utterance	-0.029	0.131***	-0.090**
	Msec. length	0.340***	-0.318***	-0.182***
	R^2	0.226	0.206	0.091
	F(6,1854)	89.99***	79.91***	30.80***

KEY: * : $p<0.01$ ** : $p<0.001$ *** : $p<0.0001$

shorter, and close to the beginning of the utterance. Missed recognitions become more and late recognitions less likely when there are fewer words and, therefore, fewer trials in the paradigm between them and the end of the utterance. Words may often have been missed because conversations were artificially truncated before all recognitions were achieved.

The role of the functor-contentive distinction may point to the characteristics of the human parsing device. The greater tendency of functors, all else being equal, to be recognised late may arise from their being constrained by subsequent as well as by prior context. For example, prepositions may be optional following particular verbs, but they may be obligatory between a particular verb and a particular noun. Similar constraints are difficult to imagine for content words. Multiple regression analyses demonstrated that content words and functors have the same relationship to the independent variables but different basic tendencies toward on-time and late recognition: contentives of a given length and position are more prone to be recognised with left context alone ($F(6,1849)=7.49$, $p=.00001$) and functors are more prone to late recognition ($F(6,1849)=7.66$, $p=.00001$).

EXPERIMENT 2

The experiment above provided the listener with little prior context. A control experiment was carried out to determine the degree to which the observed right-

context effect depended upon the fact that the utterances had been presented out of context. In Experiment 2, subjects were given all of the prior context originally available to the speakers.

Method

Materials. Six utterances were selected from two male and two female speakers from the first experiment. An audio-tape was produced for each speaker consisting of a copy of the original dialogue with the six critical utterances replaced by the gated versions of the same utterances as used in Experiment 1. The beginning and end of each gated utterance was signalled, as was the point where each gated stimulus was complete and the undoctored dialogue resumed.

Subjects and Procedure. Subjects were 48 students at the University of Edinburgh. Subjects heard all of the taped dialogue containing the gated utterances up to the end of the last gated item. They were instructed to listen carefully to all of the conversation and to write down the words contained in each gate of the gated utterances. The instructions for responding were as before.

Results

The results for Experiment 2 were analogous to those for Experiment 1 in all respects but one: the mean number of presentations to recognition for words actually recognised now fell to 1.90 ($t=2.24$, $df=249$, $p<.025$). While this represents the on-time recognition of more words, the overall distribution of case histories (Figure 1a) was statistically indistinguishable from that produced by the earlier experiment. The ratios of late to on-time recognitions also failed to differ significantly (Figure 1b), as did the distribution of late recognitions among the categories described above (Figure 2). Furthermore multiple regression equations for Experiment 2 stimuli failed to differ from equations for the same stimuli within Experiment 1. Interestingly, the detrimental effect of proximity to utterance onset is retained in this experiment, even though utterance onset was no longer the onset of left context.

The only difference between the recognition of words in contextualised and decontextualized utterances, then, was that in the former late recognitions were achieved with less rightwards information.

CONCLUSIONS

The analysis of a large sample of conversational speech allows us to say that while words are not recognized after their offsets in a majority of cases, they are recognized late often enough to demand some mechanism beyond a marginal addendum to current on-line on-time models.

We have seen that reliance on right context is greatest where acoustic ambiguity is likely to be greatest (i.e. in short word tokens); where there is less syntactic context (i.e. early in the utterance); and where the chances to reduce the set of candidates by means of longer term structural restrictions seems to be best (i.e. among functors).

The results, then, militate for the development of models of word recognition in running speech which can incorporate right context effects up to and including

those which most clearly involve right-to-left processing. It remains to be seen what levels of complication and computational expense are involved in a psychologically verifiable computational model of the phenomena described here.

ACKNOWLEDGMENTS

This work was supported by SERC Project Grant GR C78377 to the first author and H.S. Thompson. The authors wish to thank Dr. Thompson and members of the Centre for Cognitive Science Speech Processing Workshop for discussion and encouragement and Professor Gillian Brown for access to materials. Order of authorship is arbitrary.

REFERENCES

[1] W.D. Marslen-Wilson and A. Welsh. 'Processing interactions during word-recognition in continuous speech'. Cog. Psych., Vol. 10, 29-63, (1978).

[2] W.D. Marslen-Wilson and L.K. Tyler. 'The temporal structure of spoken language understanding'. Cognition, Vol. 8, 1-71, (1980).

[3] L.K. Tyler and J. Wessels, 'Quantifying contextual contributions to word-recognition processes', Percep. and Psychophys., Vol. 34, 409-420, (1983).

[4] R. A. Cole and J. Jakimik, 'A model of speech production', In R. A. Cole, (Ed.), Perception and production of fluent speech, Hillsdale, N.J.: LEA, (1980).

[5] J.M. Pickett and I. Pollack, 'Intelligibility of excerpts from fluent speech: effects of rate of utterance and duration of excerpt', Lang. and Speech, Vol. 6, 151-165, (1963).

[6] I. Pollack and J.M. Pickett, 'The intelligibility of excerpts from conversation', Lang. and Speech, Vol. 6, 165-171, (1963).

[7] F. Grosjean, 'The recognition of words after their acoustic offset: evidence and implications', Percep. and Psychophys., Vol. 38, no. 4, 299-310, (1985).

[8] J.M. McAllister, L. Wheeldon and E.G. Bard, 'What isn't in a word? – the recognition of words in read and conversational speech', unpublished paper: Department of Linguistics, University of Edinburgh. (In preparation).

[9] G. Brown, K. Currie, and J. Kenworthy., Questions of Intonation, London: Croom Helm, (1980).

[10] E.G. Bard and A.H. Anderson, 'The unintelligibility of speech to children' J. Child Lang., Vol. 10, 265-292, (1983).

[11] F. Grosjean and J. Gee, 'Prosodic structure and spoken word recognition', Cognition, (in press).

[12] W.N. Francis and H. Kucera, Frequency Analysis of English Usage, Boston: Houghton Mifflin, (1982).