Hierarchial chunking in sentence processing

W. J. M. LEVELT[1]
GRONINGEN UNIVERSITY

In order to evaluate a left-to-right hierarchical chunking model of sentence perception, Johnson's Hierarchical Clustering Scheme (HCS) technique was applied to data obtained from sentence intelligibility tests. One hundred and twenty Ss listened to sentences disturbed by white noise. After each presentation they wrote down what they had heard. For each sentence, a table of conditional probabilities p(j/i) was computed, where p(j/i) is the probability that word j had been correctly identified, given correct identification of word i. This was done for all i's and j's from the sentence. HCS analysis of the off-diagonal submatrices for which words i precede words j ("forward conditional probabilities") yielded satisfactory results. Apparently there is a latent hierarchical structure to these data. The large chunks that appear from these analyses do generally correspond to major syntactic constituents. Minor constituents, however, are very often not reflected in the chunking pattern.

Since Miller (1956) introduced the *chunk* as a unit of immediate memory, *recoding* has become a vital concept in human information processing research. With the aid of a powerful recoding system one is able to store large amounts of information into a very limited number of chunks. Miller indicated that language can be considered as a preeminent recoding system.

What are the chunks, then, in speech transmission? What can be said about the perceptual recoding of syntactic material? The obvious step is to consider linguistic units as possible candidates for chunks of speech. Miller (1962) suggests that the *phrase* might be the natural decision unit for speech. Several experiments support this view. The most striking results in this regard have been obtained by the click procedure (Fodor & Bever, 1965; Garrett, Bever, & Fodor, 1966): If a click is superimposed on a recording of continuous speech, the listener tends to dislocate the click perceptually towards a major constituent boundary.

As soon as the notion of phrase structure or constituent structure is introduced in studies of perceptual segmentation one cannot evade the question of the hierarchical order of the perceptual chunks. Constituent structure is essentially hierarchical. It suggests that decisions on the level of major constituents may be dependent upon preliminary decisions at lower levels. But also, if we take sentence understanding mainly to be a left-to-right processing, then decisions about words or small phrases may be highly dependent upon the understanding of a major preceding constituent.

Hierarchical models of sentence transmission have been proposed by several researchers. Yngve's (1960) theory is an explicit case. Osgood's (1963) theory, though less strictly a left-to-right model, is also basically a hierarchical one. Miller & Chomsky (1963) suggest that the output of a first superficial analysis of a sentence by the listener may be the surface phrase-marker.

Thus, under the influence of linguistics, rather elaborate theories have been proposed about sentence processing. Rommetveit (1968), however, rightly remarks that though there is firm experimental evidence for grouping and chunking processes, the existing information is hardly at variance with any modern theory of speech perception. In particular, there is a large gap between the intricacy of hierarchical theories, on the one hand, and the roughness of the supporting evidence, on the other hand.

It is our conviction that this gap is not due to a lack of precise data in the first place, but rather to inadequate data analysis. If one wants to test a hierarchical chunking theory of sentence perception, methods of analysis should be used that are explicitly designed for the assessment of hierarchical structures.

The present study was conducted to create the possibility of revealing a potential latent hierarchical structure in data on sentence perception.

METHOD

The experiment was essentially an intelligibility test. Spoken sentences embedded in white noise were presented to Ss. At each presentation they had to write down what they heard.

Stimulus Material

Twenty Dutch sentences were composed of various syntactic structures. We were careful to introduce three sentences of the type *the tall boy saved a dying woman*, with two major constituents, as well as three sentences of the type *the house across the street was burning*, which has three major constituents. These syntactic types have profitably been used by N. F. Johnson (1965) in a study on transitional errors in sentence recall. We used several sentences with direct and indirect objects like *the boy gave the ice cream to a child* and added passive and question versions of such syntactic structures. The sentences that were submitted to further analysis are given in Tables 1-11, with their word-to-word English translations. Care was taken that in each sentence all words were different.

The 20 sentences were spoken by an adult male voice[2] and tape-recorded in a sound proof room by means of high-quality recording equipment (Sennheiser microphone, Revox A77 recorder). The order of the sentences was random, except that immediate succession of sentences with the same syntactic structure was prevented.

Next, white noise was recorded on the second tape track. In a preexperiment, a S/N ratio was determined that yielded about 50% correct identification of all the words on the tape. The mixed signal was presented via a loudspeaker.

Subjects

In the main experiment, Ss were 120 undergraduate psychology students, both men and women.

Procedure

The noise-embedded sentences were presented one by one via a loudspeaker. The Ss were instructed to listen carefully and to write down after each presentation what they had heard. They were provided with test booklets with a page for each sentence. They were not allowed to go back and make changes on earlier pages.

Scoring

For each S and sentence it was determined which words had been correctly identified. We found a total of 40% correct identifications. However, the intelligibility level varied widely for the different sentences, with extremes of 11% and 87%. For the purpose of the further analysis we could only use the middle range from 30% to 70%. This excluded nine sentences.

Table 1
Conditional Probabilities for *De grote jongen redde een stervende vrouw*

	The	tall	boy	saved	a	dying	woman
1	1.000	.771	.477	.055	.028	.046	.064
2	1.000	1.000	.595	.060	.036	.024	.071
3	1.000	.962	1.000	.096	.058	.058	.115
4	1.000	.833	.833	1.000	.333	.333	.167
5	1.000	1.000	1.000	.667	1.000	.333	.333
6	.833	.333	.500	.333	.167	1.000	.167
7	1.000	.857	.857	.143	.143	.143	1.000

Table 2
Conditional Probabilities for *Het kind van de buren komt op tijd*

	The	child	of	the	neighbors	comes	in	time
1	1.000	.821	.755	.736	.679	.094	.208	.123
2	.978	1.000	.764	.764	.742	.090	.236	.135
3	.952	.810	1.000	.964	.845	.119	.190	.095
4	.963	.840	1.000	1.000	.877	.123	.198	.099
5	.986	.904	.973	.973	1.000	.137	.233	.110
6	1.000	.800	1.000	1.000	1.000	1.000	.600	.600
7	1.000	.955	.727	.727	.773	.273	1.000	.500
8	1.000	.923	.615	.615	.615	.462	.846	1.000

Table 3
Conditional Probabilities for *Het water onder de brug draait in kolken*

	The	water	under	the	bridge	whirls	in	eddies
1	1.000	.819	.298	.394	.372	.106	.096	.128
2	.987	1.000	.372	.487	.462	.141	.115	.154
3	.933	.967	1.000	.967	.933	.267	.267	.300
4	.925	.950	.725	1.000	.850	.200	.200	.225
5	.875	.900	.700	.850	1.000	.225	.175	.225
6	.909	1.000	.727	.727	.818	1.000	.727	.636
7	1.000	1.000	.889	.889	.778	.889	1.000	.778
8	1.000	1.000	.750	.750	.750	.583	.583	1.000

Table 4
Conditional Probabilities for *Het huis van de bakker staat in brand*

	The	house	of	the	baker	is	on	fire
1	1.000	.983	.746	.720	.610	.475	.381	.331
2	1.000	1.000	.759	.733	.621	.483	.388	.336
3	1.000	1.000	1.000	.920	.807	.614	.477	.409
4	1.000	1.000	.953	1.000	.824	.624	.482	.412
5	1.000	1.000	.986	.972	1.000	.694	.528	.444
6	1.000	1.000	.964	.946	.893	1.000	.786	.679
7	1.000	1.000	.933	.911	.844	.978	1.000	.844
8	1.000	1.000	.923	.897	.821	.974	.974	1.000

Table 5
Conditional Probabilities for *De directeur stuurde het honorarium aan hem*

	The	director	sent	the	fee	to	him
1	1.000	.957	.914	.741	.466	.612	.440
2	1.000	1.000	.937	.775	.486	.622	.450
3	1.000	.981	1.000	.802	.500	.660	.472
4	1.000	1.000	.988	1.000	.616	.744	.535
5	1.000	1.000	.981	.981	1.000	.852	.667
6	1.000	.972	.986	.901	.648	1.000	.704
7	1.000	.980	.980	.902	.706	.980	1.000

Table 6
Conditional Probabilities for *De schuur van het boerderijtje valt in puin*

	The	barn	of	the	farm	falls	in	ruins
1	1.000	.680	.738	.612	.699	.350	.553	.534
2	1.000	1.000	.857	.786	.871	.443	.686	.671
3	.962	.759	1.000	.810	.823	.430	.608	.608
4	.969	.846	.985	1.000	.862	.492	.692	.692
5	.973	.824	.878	.757	1.000	.432	.649	.649
6	.947	.816	.895	.842	.842	1.000	1.000	.974
7	.950	.800	.800	.750	.800	.633	1.000	.950
8	.965	.825	.842	.789	.842	.649	1.000	1.000

Conditional Probabilities

For the remaining 11 sentences, we computed tables of conditional probabilities.[3] If i and j are words from a given sentence, the material allowed for the determination of p(j/i) and p(i/j). These are, respectively, the probability that j is correctly identified, given correct identification of i, and inversely. If i and j have this order in the sentence, it is convenient to say that p(j/i) is a *forward conditional probability* and p(i/j) a *backward conditional probability*. For each sentence we computed these probabilities for all i and j from the sentence. They are given in Tables 1 through 11.

Hierarchical Clustering Scheme Analysis

In order to investigate whether or not there is a latent hierarchical structure underlying these data, we applied S. Johnson's (1967) Hierarchical Clustering Scheme (HCS) analysis. In essence, this is an algorithm that maps relatedness data onto a tree structure. Starting from n objects and their relatedness values, an iterative procedure merges objects into clusters and clusters and/or objects into larger clusters. Each new clustering is obtained by merging clusters and/or objects at the previous level. At the final level, all objects are in one cluster. An advantage of Johnson's method is that the order of the clusters in the hierarchy is insensitive to monotonic transformation of the relatedness values. A clustering value is assigned to each cluster in the tree (to each *node*, one could say). A clustering value is a measure for the "strength" of the association between the objects in the cluster. If the requirement is made that the clustering is invariant under monotonic transformations of the relatedness data, there are two ways to assign values to the clusters. The first way is to take the *smallest* relatedness value between the objects within the cluster as a measure for the strength of a cluster. The cluster value, then, indicates that all relatedness values between the elements of the cluster are larger than or equal to this value. This is called the *diameter method*, because the clustering algorithm attempts at each stage to minimize the *diameter* of the cluster. The diameter of the cluster is the largest intracluster distance (or, in terms of relatedness: the smallest intracluster relatedness value). The second method is called the *connectedness method*. In this case, the clustering value means the following: If we take any pair of objects i, j from the cluster, it is always possible to find a chain of objects from the cluster, starting at i and ending at j, such that all adjacent objects in this chain have a relatedness value that is larger than or

equal to the clustering value. The connectedness method, therefore, attempts to make clusters such that one can always "get" from i to j within a cluster via a series of "steps" that are as small as possible.

At this point it should be remarked that, theoretically, application of the connectedness method is not justified if relatedness values are asymmetrical. This is the case for conditional word identification probabilities. To get from i to j within a cluster via a chain of minimal steps is rather meaningless if the chain does not correspond to the word order in the sentence. But the connectedness method pays no regard to the direction of the relatedness values.

The asymmetry of the conditional probabilities is, however, immaterial for the diameter method. The further analysis will therefore be based on the diameter method. That the connectedness method is nevertheless included is due to the fact that if the structure of the data is completely hierarchical *all* chains from i to j in a cluster have the same maximal step size. In the ideal case, therefore, the above argument is vacuous.

Johnson showed that one can define a distance metric d for such hierarchical clustering schemes. The metric is stronger than a Euclidian distance metric. It not only satisfies the triangular inequality, but also the so-called *ultrametric inequality*, namely $d(x, z) \leqslant \max [d(x, y), d(y, z)]$. For further details we must refer to Johnson's original article.

Table 7
Conditional Probabilities for *De nieuwe auto ramde een betonnen paal*

	The	new	car	rammed	a	concrete	pale
1	1.000	.811	.793	.351	.243	.117	.315
2	.989	1.000	.923	.396	.275	.143	.363
3	.989	.944	1.000	.404	.281	.146	.348
4	.975	.900	.900	1.000	.675	.325	.750
5	1.000	.926	.926	1.000	1.000	.481	.852
6	.929	.929	.929	.929	.929	1.000	1.000
7	.921	.868	.816	.789	.605	.368	1.000

Table 8
Conditional Probabilities for *De man belde op naar zijn vorige baas*

	The	man	called	up	to	his	former	boss
1	1.000	.785	.374	.561	.720	.318	.178	.196
2	.988	1.000	.447	.635	.824	.388	.200	.224
3	1.000	.950	1.000	.950	.975	.575	.350	.400
4	.968	.871	.613	1.000	.871	.435	.290	.306
5	.963	.875	.487	.675	1.000	.425	.250	.263
6	1.000	.971	.676	.794	1.000	1.000	.471	.559
7	.950	.850	.700	.900	1.000	.800	1.000	.750
8	.955	.864	.727	.864	.955	.864	.682	1.000

Table 9
Conditional Probabilities for *De oude paarden aten het malse hooi*

	The	old	horses	ate	the	tender	hay
1	1.000	.505	.421	.252	.262	.215	.374
2	1.000	1.000	.556	.333	.296	.259	.444
3	1.000	.667	1.000	.578	.578	.489	.756
4	1.000	.667	.963	1.000	.815	.704	.926
5	1.000	.571	.929	.786	1.000	.714	.964
6	1.000	.609	.957	.826	.870	1.000	1.000
7	.976	.585	.829	.610	.659	.561	1.000

Table 10
Conditional Probabilities for *De jongen gaf het ijsje aan een kind*

	The	boy	gave	the	ice cream	to	a	child
1	1.000	.852	.583	.591	.374	.678	.209	.443
2	1.000	1.000	.602	.653	.378	.684	.235	.429
3	1.000	.881	1.000	.761	.552	.821	.254	.627
4	1.000	.941	.750	1.000	.382	.735	.309	.485
5	.977	.841	.841	.591	1.000	.864	.182	.795
6	1.000	.859	.705	.641	.487	1.000	.231	.590
7	.960	.920	.680	.840	.320	.720	1.000	.520
8	.962	.792	.792	.623	.660	.868	.245	1.000

Table 11
Conditional Probabilities for *Hij betaalde het geld aan een agent*

	He	paid	the	money	to	a	police-man
1	1.000	.255	.511	.287	.500	.128	.319
2	1.000	1.000	.750	.792	.833	.167	.542
3	.980	.367	1.000	.510	.653	.122	.429
4	.964	.679	.893	1.000	.857	.179	.571
5	.959	.408	.653	.490	1.000	.245	.571
6	1.000	.333	.500	.417	1.000	1.000	.917
7	.968	.419	.677	.516	.903	.355	1.000

Goodness of Fit Measures

It can be shown that if the diameter and connectedness methods give identical results, the data do not violate the ultrametric inequality. This, then, means that there is indeed a latent hierarchical structure to the data. In this way, one can evaluate the correctness of the hierarchical assumptions.

Another criterion for the applicability of a hierarchical model is a more obvious one. Johnson's HCS, being insensitive to monotonic transformations of the relatedness data, uses only the *order* of these values. If one assigns an HCS to a set of objects, one essentially assigns (ultrametric) distance values to all pairs of objects. If the clustering is adequate, these distance values should have a monotonic inverse relation to the original relatedness data. The goodness of fit can therefore be determined from the number of order relations specified by the HCS, which are violated by the data, i.e., the conditional probabilities. This brings us to the definition of a stress measure: The *stress* of an HCS is the number of order relations violated by the data divided by the total number of order relations specified by the HCS.

Forward Conditional Probabilities

The right upper halves of Tables 1-11 were subjected to diameter and connectedness HCS analyses. We computed the stress values for all HCSs obtained. They are given in Table 12.

The table shows that on the average the diameter HCS solutions violate only 5.3% of the rank orderings in the data matrices. In view of the strong limitations that must be satisfied by a data matrix for the ultrametric inequality to hold, one is inclined to take this low number of violations of order as a confirmation of a latent hierarchical structure in the data matrices. A closer look at the table,

Table 12
Stress Values for HCS-Solutions

Sentence No.	1	2	3	4	5	6	7	8	9	10	11	Mean
Diameter Method	0	0	0.4	1.2	1.4	5.0	5.8	7.0	10.2	13.0	14.1	5.3%
Connectedness Method	0	0	0.4	0	11.7	4.6	6.5	11.0	9.7	15.0	13.4	6.6%

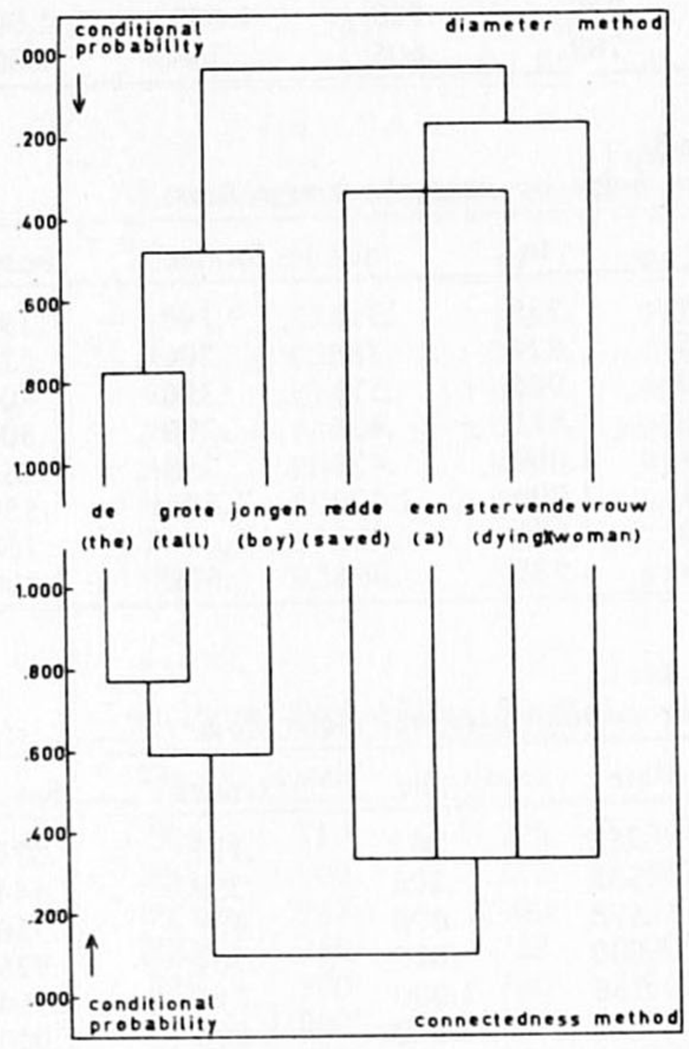

Fig. 1. HCS solution (diameter method, upper half; connectedness method, lower half) for "De grote jongen redde een stervende vrouw [The tall boy saved a dying woman]." (Forward conditional probability data.)

however, asks for some qualification of this general conclusion. Diameter stress values vary from 0% to 14.1%, and one can expect a larger range if a larger sample of sentences is used. Hence we do not want to draw general conclusions from this limited set of experimental data.

On the other hand, many, in fact the first 5 of these 11 sentences, show amazingly little stress. It seems to be worthwhile to give them a more detailed inspection.

Figures 1 through 5 give the hierarchical clusterings that were obtained for these sentences. The upper half of each figure represents the diameter method clustering, the lower half the connectedness clustering.

In all cases, the two methods give virtually identical hierarchies. This is another indication for the latent hierarchical structure in the corresponding data matrices.

The higher the left-to-right path from one word to another in these trees, the lower the influence of the earlier word on the intelligibility of the later one. It is therefore justified to interpret these trees as adequate representations of the

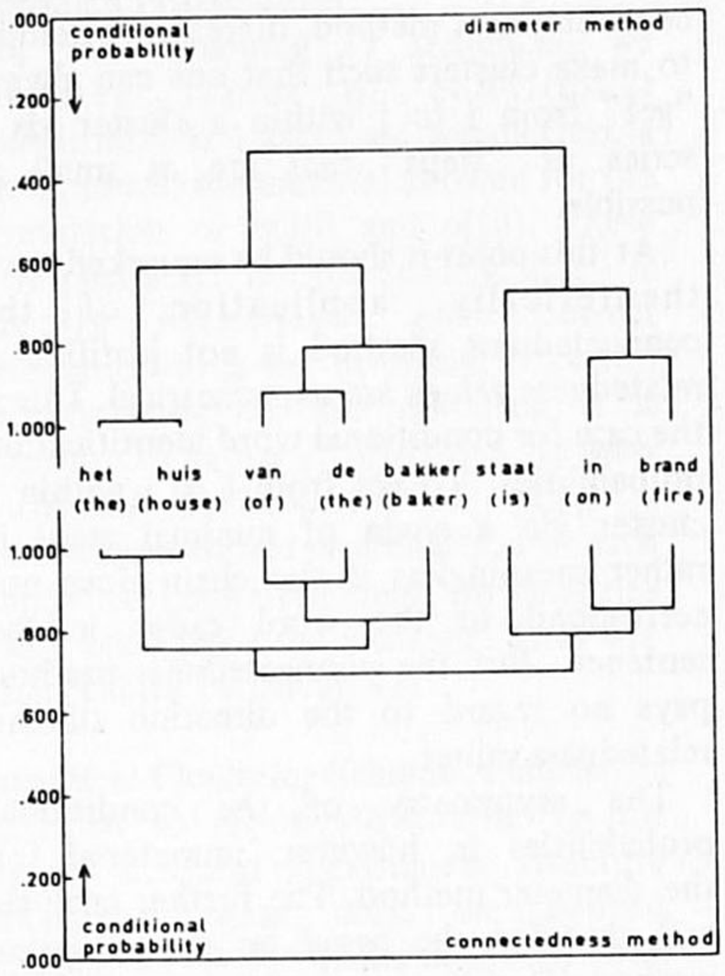

Fig. 4. HCS solution (diameter method, upper half; connectedness method, lower half) for "Het huis van de bakker staat in brand [The house of the baker is on fire]." (Forward conditional probability data.)

left-to-right chunking that takes place in the processing of these sentences. There are several interesting aspects to these figures: (1) None of the trees shows any crossing of lines; in all cases chunks consist of *adjacent* words or word groupings. This is *not* an artifact of the clustering technique. (2) In general, large chunks correspond to major constituents. The only exception is in

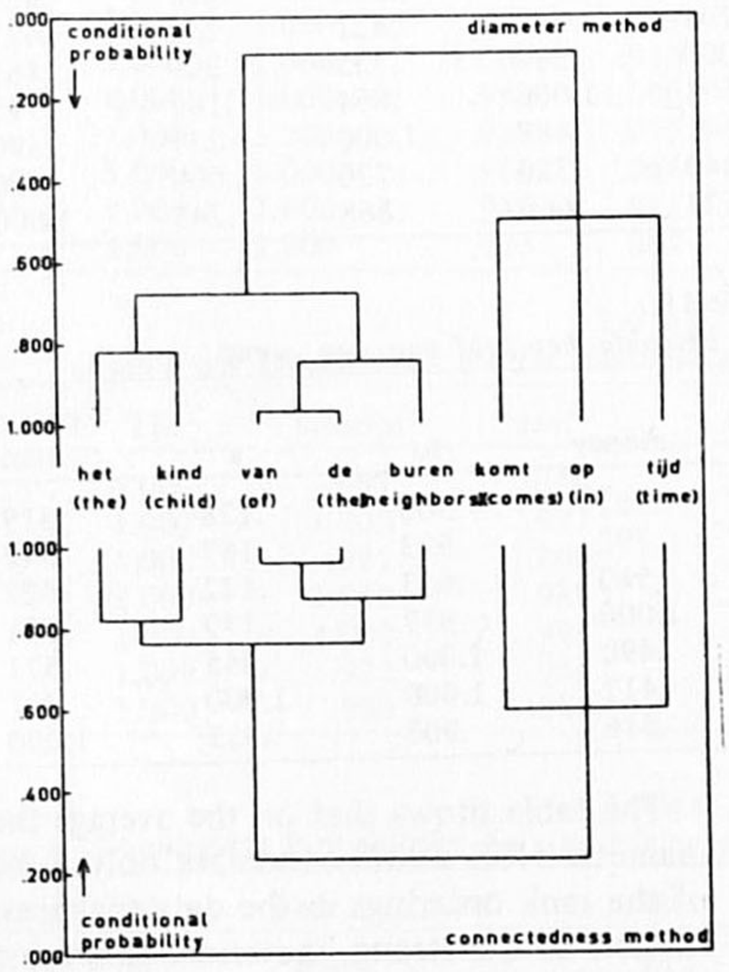

Fig. 2. HCS solution (diameter method, upper half; connectedness method, lower half) for "Het kind van de buren komt op tijd [The child of the neighbors comes in time]." (Forward conditional probability data.)

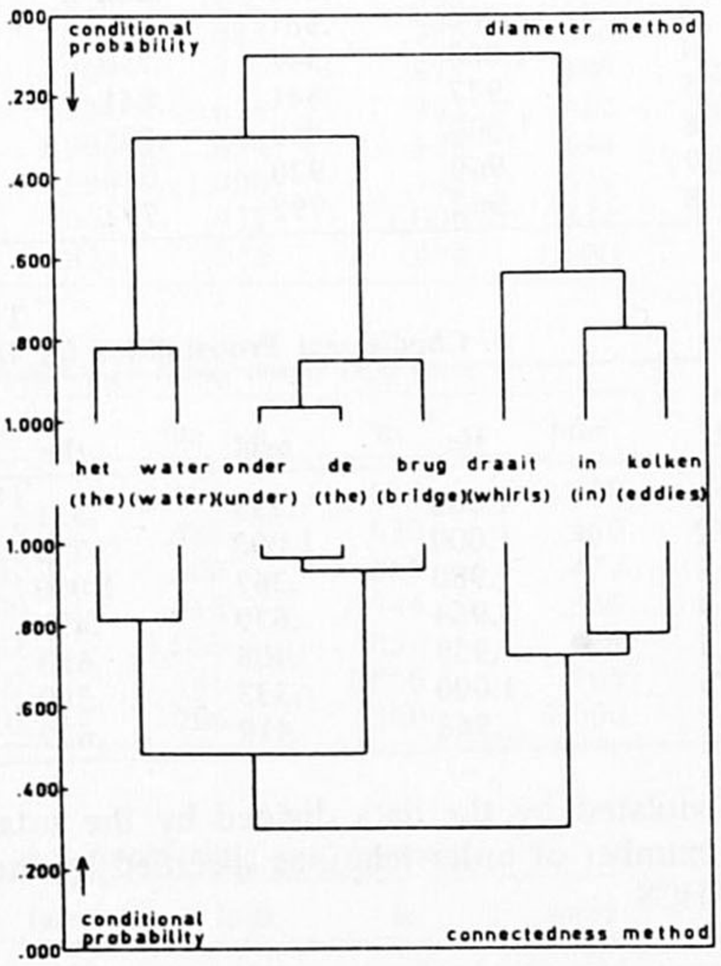

Fig. 3. HCS solution (diameter method, upper half; connectedness method, lower half) for "Het water onder de brug draait in kolken [The water under the bridge whirls in eddies]." (Forward conditional probability data.)

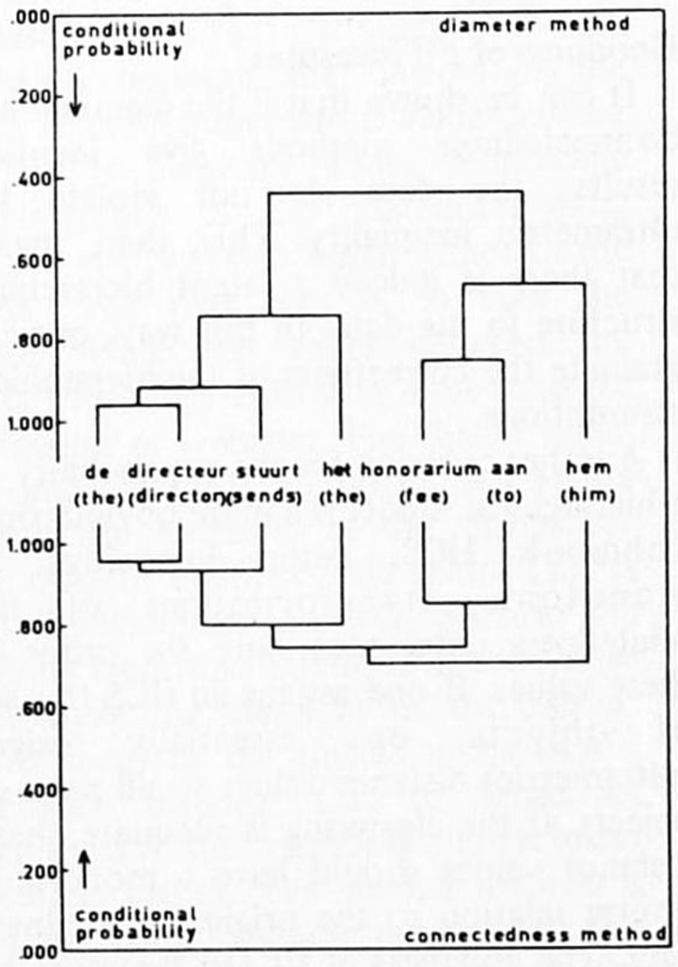

Fig. 5. HCS solution (diameter method, upper half; connectedness method, lower half) for "De directeur stuurt het honorarium aan hem [The director sends the fee to him]." (Forward conditional probability data.)

Fig. 5, where a major break occurs between *the* and *fee*. (3) Small chunks are not systematically related to the minor constituents. In particular, the article is often dissociated from its noun; there are chunks like *the tall, of the, under the.* This behavior of the article may at the same time account for the exception under (2). (4) It turns out that three of the five sentences are of the same constituent structure: *the child of the neighbors comes in time, the water under the bridge whirls in eddies* and *the house of the baker is on fire.* In fact, there were no other sentences of this type in the sample. These three sentences apparently produce virtually identical chunking patterns. This suggests that, although chunking is not fully related to surface constituent structure, sentences with identical structure give rise to identical chunking patterns.

Backward Conditional Probabilities

We can be short about the lower halves of Tables 1 through 5. The numbers are too high for a profitable application of HCS analysis: In general, one obtains one big cluster of all words, especially by the connectedness method. Undoubtedly, backwards disambiguation exists in sentence perception .The perception of a later word can a posteriori facilitate the identification of an earlier word. But in the present data not much structure is apparent in this "backwards information flow." It is incomparable to the subtle patterns we find in the forwards spread of information.

CONCLUSIONS

This study has served a triple purpose. The first objective was to find a means of data analysis that was adequate to the hierarchical structure of certain theories of sentence processing. It is shown that Johnson's HCS analysis, if applied to conditional word identification probabilities, fits this purpose. This is especially the case for the diameter method.

The second objective was to test the adequacy of hierarchical chunking models of speech. Two criteria were proposed for the goodness of fit of an HCS. The most direct one is the amount of stress of an HCS solution. This is the percentage of order relations predicted by the HCS solution that are violated by the data. For the diameter method we found an average stress value of 5.3%. This can be taken as a confirmation of the existence of a latent hierarchical structure in the forward conditional probabilities. Five of the 11 sentences were especially low in their stress values ($< 1.5\%$). These also show a remarkable accordance with the second criterion, namely the virtual identity of their diameter and connectedness solutions (Figs. 1-5). At this point the tentative conclusion is that hierarchical left-to-right chunking will often be an adequate model for sentence processing. It should be added, however, that a model for the partitioning of the syntactic input is by no means a complete model of sentence understanding. Semantically important relations like subject of the sentence, direct object, etc., are often not deducible from *any* pattern of chunks. Nevertheless, they should be discerned if a full understanding of the sentence is required. We refer to Miller & Chomsky (1963, p. 476) for a more detailed discussion of this issue.

Thirdly, if a hierarchical model is adequate, how is the chunking hierarchy related to the constituent structure of the sentence? The analyses gave rise to three tentative statements: (1) Large chunks tend to coincide with major constituents. (2) The minor constituents are not systematically reflected in the structure of small chunks. (3) Sentences with equal constituent structures are chunked in the same way.

It has not been the purpose of this paper to study the various cues that may trigger decisions in the processing of syntactic material. The question is important, however, as to how much of the chunk structure can be accounted for in terms of a "passive" recognition device, i.e., a perceptual mechanism that merely reflects the acoustical structure of the input sentence. Especially prosodic features like intonation and pause pattern may be material in making preliminary decisions on word grouping. For a further study of the role of such cues in the understanding of structurally ambiguous sentences, see Levelt, Zwanenburg, & Ouweneel (in press). But it is also known that there is active use of grammatical knowledge on the part of the listener in structuring syntactic material. In this way, the hearer is less dependent on the sound spectrum of the input sentence. He may, for instance, make conclusions about several aspects of the grammatical organization of the sentence on the basis of one or two key words he happened to recognize.

The present experimental procedure can yield information about the chunk patterns of the listener. Further systematic variation of cues will, hopefully, reveal more about the determinants of such patterns.

REFERENCES

FODOR, T. A., & BEVER, T. G. The psychological reality of linguistic segments. Journal of Verbal Learning & Verbal Behavior, 1965, 4, 414-420.

GARRETT, M., BEVER, T. G., & FODOR, J. A. The active use of grammar in speech perception. Perception & Psychophysics, 1966, 1, 30-32.

JOHNSON, N. F. The psychological reality of phrase structure rules. Journal of Verbal Learning & Verbal Behavior, 1965, 4, 469-475.

JOHNSON, S. C. Hierarchical clustering schemes. Psychometrika, 1967, 32, 241-254.

LEVELT, W. J. M., ZWANENBURG, W., & OUWENEEL, G. R. E. Ambiguous surface structure and phonetic form in French. Foundations of Language, in press.

MILLER, G. A. The magical number seven plus or minus two. Psychological Review, 1956, 63, 81-97.

MILLER, G. A. Decision units in the perception of speech. IRE Transactions on Information Theory, 1962, IT- 8, 2, 81-83.

MILLER, G. A., & CHOMSKY, N. Finitary models of language users. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology.* New York: Wiley, 1963. Pp. 419-491.

OSGOOD, C. E. On understanding and creating sentences. American Psychologist, 1963, 18, 735-751.

ROMMETVEIT, R. *Words, meaning and messages.* New York: Academic Press, 1968.

YNGVE, V. H. A model and a hypothesis for language structure. Proceedings of the American Philosophical Society, 1960, 104, 444-466.

NOTES

1. Address: Instituut voor Algemene Psychologie. Oude Boteringestraat 34, Groningen, The Netherlands.

2. Sentences were spoken by Mr. J. v.d. Sman, who also–in cooperation with Mr. H. Kobus–assisted in the computations.

3. These were, of course, conditional relative frequences, but they are taken as conditional probability estimates.

(Accepted for publication November 15, 1969.)