CHAPTER 4

Scaling procedures in the psychological study of grammar

Introduction - Hierarchical clustering algorithms in the psychology of grammar

WILLEM J. M. LEVELT

Groningen University

The papers in this chapter are both psychological studies of grammar, but they share little in terms of content. One deals with syntactical, the other one with phonological problems. The reason for nevertheless adding a common introduction is due to the method of data analysis that is used in these articles. They are quite similar in this respect but both contributions are very summary in explaining the scaling methods that are involved. The following is a short non-technical review of the relevant clustering procedures. For more formal introductions we must refer to other sources: to JOHNSON (1967) for the hierarchical clustering schemes and to LANCE and WILLIAMS (1966/1967) for a more-general review of clustering algorithms.

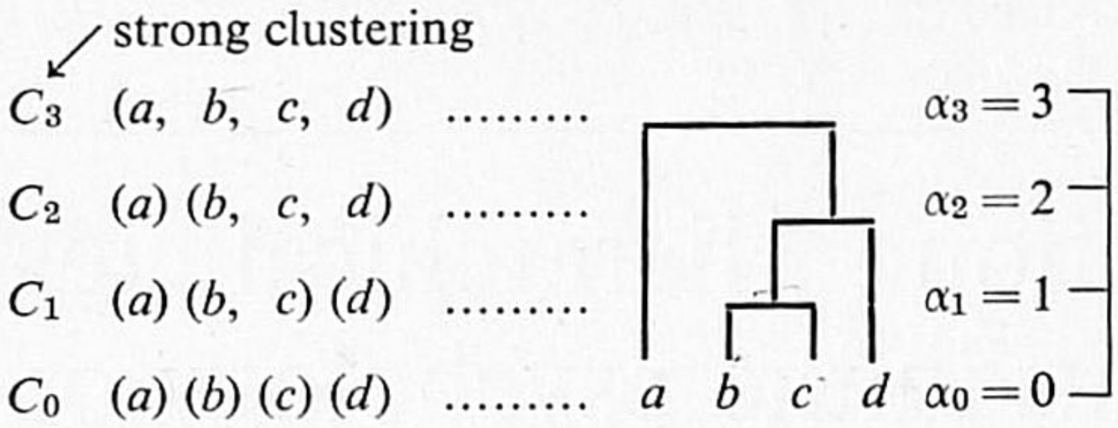
One preliminary remark: scaling methods concern both procedures for the collection of data and for analyzing data. With respect to the first aspect, data gathering, the papers in this chapter speak for themselves. In Levelt's paper the data are obtained by either triadic comparisons or by direct judgments of relatedness between words from a sentence. Campbell's data are confusions of speech sounds, essentially a-symmetrical data.

In all cases, however, the data are translated into measures of relatedness or similarity between stimuli (words, speech sounds). It is at this point that the papers converge: the analysis of similarity data.

We thus start at the situation where one has the disposal of $n(n-1)/2$ relatedness or similarity measures $s(i,j)$, one for each pair i,j $(i \neq j)$ from a set of n stimuli (words, sounds).

The aim is to find a set of clusterings of the n objects on the basis of these similarity or relatedness measures. A clustering of n objects is defined as any partitioning of all n objects in non-overlapping subgroups. There are two

Table 1. A hierarchical clustering scheme with tree and bracketing representation.



limiting cases: the *weak* clustering C_0 is the clustering consisting of n clusters of one object each, the *strong* clustering C_m is the case where all n objects are in the same cluster. An example, where $n=4$ is given in Table 1. The objects are labeled a through d. The particular set of clusterings should be hierarchical. A hierarchical clustering scheme is an ordered set of $m+1$ clusterings, starting at the weak clustering C_0 and ending at the strong clustering C_m, such that each clustering C_i $(i=1, 2, \ldots m)$ is obtained by a merging of clusters in the foregoing clustering C_{i-1}. The example in Table 1 is a hierarchical clustering scheme, where $m=3$. A hierarchical clustering scheme can as well be represented by a tree graph. The equivalent tree graph for the example is also given in Table 1.

The next step is to define a distance metric. This is done by assigning values $\alpha_0, \ldots \alpha_m$ to the respective clusterings. If the value of the weak clustering, α_0 is put to 0, whereas the other values α_i are monotonically increasing with i, the distance between two elements or objects x and y, $d(x, y)$ can easily be defined as follows: given that C_k is the weakest clustering where x and y are in the same cluster, then $d(x, y)=\alpha_k$. In Fig. 1 we have arbitrarily assigned α's to the four clusterings. For the sake of simplicity we choose: $\alpha_0=0$, $\alpha_1=1$, $\alpha_2=2$ and $\alpha_3=3$. This choice fully determines the distances between the objects 1 through 4. For instance: objects b and c are clustered at C_1, therefore $d(b, c)=\alpha_1=1$. Similarly $d(b, d)=\alpha_2=2$, $d(a, b)=\alpha_3=3$. The distance matrix D_1 corresponding with Table 1 is given in Table 2. Without further proof it should be clear that d has all the properties for it to be a distance metric: (1) $d(x, y)=d(y, x)$ (x and y are clustered at the same level as y and x), (2) $d(x, x)=0$ (all elements are clustered with themselves in the weak clustering, where $\alpha=0$). (3) The triangular inequality holds: $d(x, z) \leqslant d(x, y)+d(y, z)$. In fact a much stronger inequality, the so-called "ultrametric inequality" is valid: $d(x, z) \leqslant \max[d(x, y), d(y, z)]$. A distance between

Table 2. Layout of hierarchical clustering algorithm.

$D_1 \longrightarrow D_2 \longrightarrow D_3 \longrightarrow D_4$

	a	b	c	d
a	0	3	3	3
b	3	0	**1**	2
c	3	**1**	0	2
d	3	2	2	0

	a	(b, c)	d
a	0	3	3
(b, c)	3	0	**2**
d	3	**2**	0

	a	$((b, c)\, d)$
a	0	**3**
$((b, c)\, d)$	**3**	0

	$(a((b, c)\, d))$
$(a((b, c)\, d))$	0

two objects is never larger than the largest distance of these objects to a common third object.[1]

Given the distance matrix D_1 in Table 2 it is possible to reconstruct the clustering scheme in Table 1. Johnson proposed the following procedure:

1) Find the smallest distance >0 in D_1. For the example $d(b, c)=1$ (bold type in Table 2).
2) Make a cluster of the corresponding objects x and y (b and c) and determine the distances of this cluster to the other objects $d([x, y], i)$, i.e. $d([b, c], a)$ and $d([b, c], d)$. For the example, it can be easily verified that $d(b, a)=d(c, a)$ and $d(b, d)=d(c, d)$. It can be proved that $d(x, i)=$ $=d(y, i)$ is generally true if x and y are in the same cluster and if the ultrametric inequality holds. The natural definition of the distance between a cluster (x, y) and other clusters or objects, i, is therefore:
   (1) $d([x, y], i)=d(x, i)=d(y, i)$.
   In the present case $d([b, c], a)=d(b, a)=d(c, a)=3$ and $d([b, c], d)=$ $=d(b, d)=d(c, d)=2$. This clustering of objects results in a new distance matrix D_2, also given in Table 2.
3) Repeat the procedure until the strong clustering has been obtained – see Table 2. It is clear from this table that the original hierarchical clustering scheme has been recovered.

There are, however, two points that need further consideration if we want to analyze real data, i.e. experimentally obtained similarity or relatedness data.

1. Where the data are similarities, not distances, the algorithm, exemplified in Table 2, should be "translated". On the assumption, however, that similarities are inversely related to distances the adaptation can be simple. Starting from the similarity matrix S_1, the first clustering is based on the

[1] This, actually, is the state of affairs for equilateral triangles with vertical angle $\leqslant 60°$.

largest similarity value $s(x, y)$ $(x \neq y)$ in S_1. New similarities between the cluster (x, y) and other objects are determined in the same way as it was done for distances in the original algorithm, i.e. $s([x, y], i) = s(x, i) = s(y, i)$ for $i \neq x, y$. The further iterative process is obvious.

2. Due to experimental error or to violation of the assumed ultrametric data structure equality (1) will in general not obtain. In term of similarities: if $s(x, y)$ is the maximal similarity in the matrix, it will in general not be the case that $s(x, i) = s(y, i)$ for all $i \neq x, y$. If it is decided to cluster x and y because of their maximal similarity, some ad hoc decision has therefore to be made in order to define the similarity between the new cluster and the remaining objects or clusters, $s([x, y], i)$. Lance and Williams summarize various procedures to this effect. Two of them are of particular interest; they are the ones used by Johnson and also in the papers of this chapter:

(a) *Connectedness method.* In this procedure $s([x, y], i)$ is defined as $\max[s(x, i), s(y, i)]$ i.e. in merging the rows and columns of the objects in S_1 that are to be clustered, one replaces each pair of conjoint similarities by the largest one. Table 3 exemplifies the application of the connectedness method to a 4×4 similarity matrix. In S_1 the maximum similarity is $s(c, d) = 6$. It is therefore decided to cluster objects c and d. The new similarity $s([c, d], a)$ is taken to be $\max[s(c, a) = 1, \; s(d, a) = 2] = 2$. Similarly $s([c, d], b) = = \max[s(c, b), \; s(d, b)] = 5$. Iteration of this procedure continues until the strong clustering is obtained.

(b) *Diameter method.* This procedure differs from the connectedness method only in that $s([x, y], i)$ is taken to be $\min[s(x, i), s(y, i)]$: The choice of the pair of objects to be clustered is, as in the connectedness method, solely based on the maximum similarity value in S, but in the process of merging, each pair of conjoint similarities is replaced by the smallest member. This is also exemplified in Table 3.

The example shows that the two procedures may yield a different result. This is immediately apparent from the tree graphs in the table. They are topologically different, as well as numerically. Notice that the numbers at the vertical axis correspond to the bold type numbers in the matrices, i.e. the similarity values on which the clustering decisions have been based. At this point three questions arise: First, why two methods with different solution, instead of one, based on some averaging procedure? Second, can one find theoretical arguments for making a particular choice between the two methods? And finally: how do we evaluate differences in solution obtained by the two methods?

Table 3. Connectedness and diameter methods, applied to a similarity matrix. Tree representations and "model" ultrametric similarity matrices.

S_1

	a	b	c	d
a	–	4	1	2
b	4	–	3	5
c	1	3	–	**6**
d	2	5	**6**	–

Connectedness method

Diameter method

S_2

	a	b	(c, d)
a	–	4	2
b	4	–	**5**
(c, d)	2	**5**	–

	a	b	(c, d)
a	–	**4**	1
b	**4**	–	3
(c, d)	1	3	–

S_3

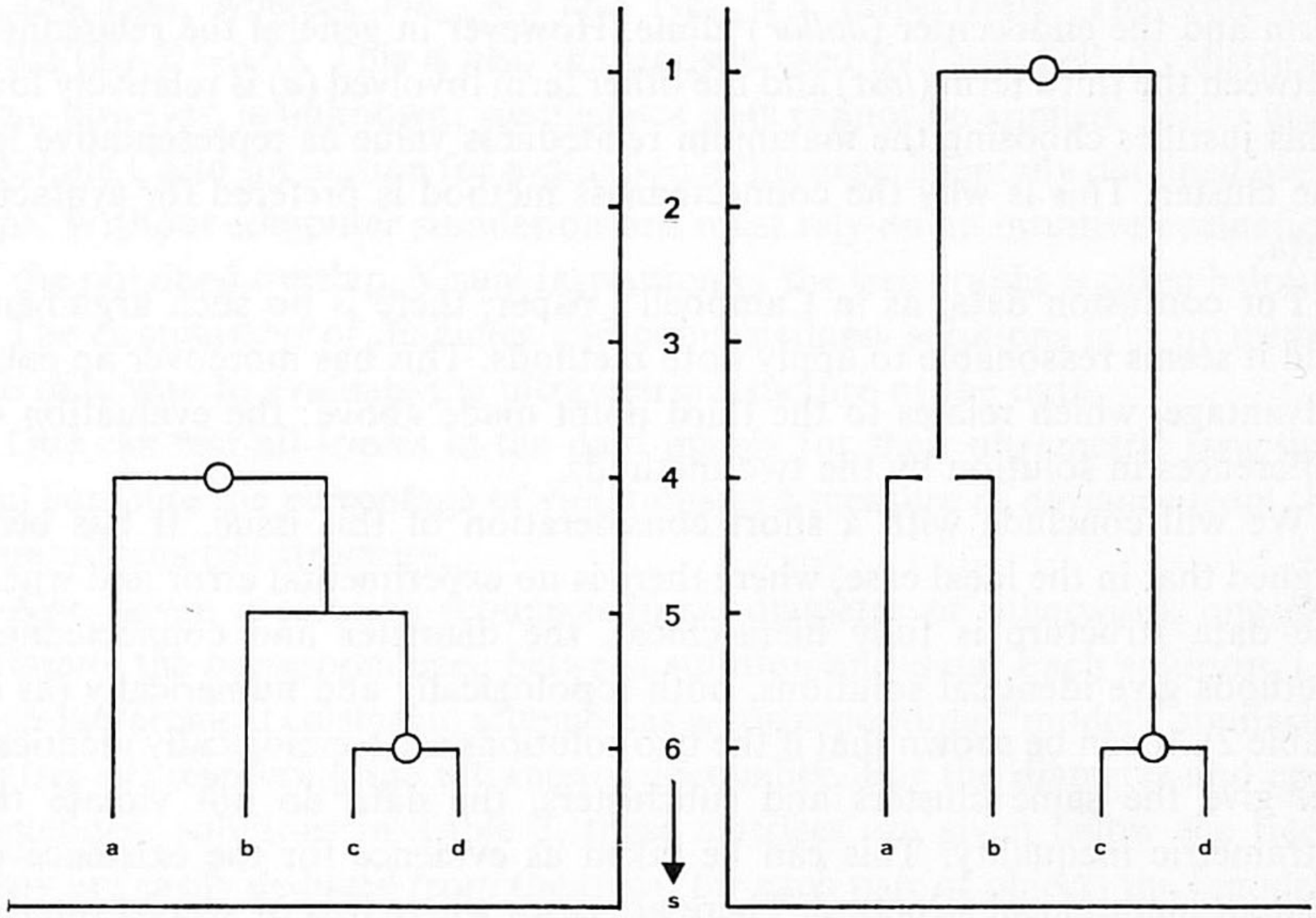
	a	$(b(c, d))$
a	–	**4**
$(b(c, d))$	**4**	–

	(a, b)	(c, d)
(a, b)	–	**1**
(c, d)	**1**	–

S_4

	$(a(b(c, d)))$
$(a(b(c, d)))$	–

	$((a, b)\ (c, d))$
$((a, b)\ (c, d))$	–



M

	a	b	c	d
a	–	4	4	4
b	4	–	5	5
c	4	5	–	6
d	4	5	6	–

	a	b	c	d
a	–	4	1	1
b	4	–	1	1
c	1	1	–	6
d	1	1	6	–

The first question is easily answered. Any algorithm where the similarity $s([x, y], i)$ is computed from $s(x, i)$ and $s(y, i)$ by an averaging procedure will not yield solutions that are invariant under monotone transformations of the similarity matrix. Diameter and connectedness methods give invariant solutions.

With respect to the second point, which of the two methods is more appropriate for a given set of data, the answer, of course, depends on the character of the data. There may be reasons to assume that $s([x, y], i)$ is better approximated by the maximum of $s(x, i)$ and $s(y, i)$ in certain cases. For Levelt's data this seems to be a reasonable assumption. His data are relatedness estimations for all pairs of words from a test sentence. Take his sentence *the boy has lost a dollar*. If the data indicate that *a* and *dollar* should form a cluster, the relatedness values between *a dollar* and the other objects have to be estimated. For *s*(*lost*, *a dollar*) the choice is between *s*(*lost*, *a*) and *s*(*lost*, *dollar*). Intuitively the latter is a much better estimation than the first one. This may be due to the fact that *dollar* is the endocenter of the word group *a dollar*. Empirical tests (as yet unpublished) show that in such cases there is indeed no significant difference between subjects' estimations of on the one hand the relatedness between a third term (i.e. *lost*) and the word group (*a dollar*) as a whole, and on the other hand between the third term and the endocenter (*dollar*) alone. However in general the relatedness between the third term (*lost*) and the other term involved (*a*) is relatively low. This justifies choosing the maximum relatedness value as representative for the cluster. This is why the connectedness method is prefered for syntactic data.

For confusion data, as in Campbell's paper, there is no such argument, and it seems reasonable to apply both methods. This has moreover an extra advantage, which relates to the third point made above: the evaluation of differences in solution by the two methods.

We will conclude with a short consideration of this issue. It has been argued that in the ideal case, where there is no experimental error and where the data structure is fully hierarchical, the diameter and connectedness methods give identical solutions, both topologically and numerically (as in Table 2). It can be shown that if the two solutions are topologically identical, i.e. give the same clusters and subclusters, the data do not violate the ultrametric inequality. This can be taken as evidence for the existence of a latent hierarchical structure. There are cases where it is of special interest to test such an hypothesis. In Campbell's paper, for instance, the central issue is whether phonetic features are processed successively. A hierarchy

of features should be reflected in the confusion data. The obvious procedure, then, is to apply both algorithms to the similarity matrix, and to compare the solutions. If they are identical one may conclude that the structure of the data is ultrametric, i.e. hierarchical. It must be added however, that the situation is slightly more complicated. It is necessary to define a statistic for evaluating the correspondence of the two solutions. There are various ways to do this. MILLER (1969) counted the number of common clusters in the two solutions (excluding the weak clustering: single objects are not counted) and devides this by the average number of clusters in the two solutions, an overlap measure, one could say. In formula:

$$0 = \frac{2Nc_c}{Nc_{dia} + Nc_{con}},$$

where Nc_{dia} and Nc_{con} are the number of clusters in the diameter and connectedness solutions and Nc_c is the number of clusters they have in common. In the example in Table 3 we have indicated clusters that are common for the two solutions by open circles in the trees. It is easily seen that $Nc_c = 2$ in this case, whereas $Nc_{con} = 3$ and $Nc_{dia} = 3$, respectively. Therefore $0 = 2.2/(3+3) = 0.67$. This is also the statistic used by Campbell. 0's distribution, however, is unknown; significance tests cannot be applied. Miller used a Monte Carlo simulation for evaluation of his experimentally obtained overlaps. Without computer simulation one must rely on an intuitive evaluation of the obtained overlap. Visual inspection of the tree graphs is often helpful.

The comparison of diameter and connectedness solutions is by no means the only way to evaluate the ultrametric structure of the data.

One can test all triples in the data matrix for their ultrametric structure and compute the percentage of violations as a measure of deviance from the pure ultrametric structure.

Also, given a solution (connectedness, diameter or otherwise), one can compute the correspondence between solution and data. Each solution, i.e. each hierarchical clustering scheme has a corresponding "model" similarity matrix M, respecting the ultrametric inequality. For the diameter and connectedness solutions in Table 3, these matrices are given below the trees. They are easily deduced from the trees: for each pair of objects the "model" similarity corresponds to the value of the lowest node in the tree connecting the objects. Any solution can be evaluated by computing a cell by cell measure of association between data matrix and model matrix (for instance a correlation coefficient). Also one could compute a measure of noncorre-

spondence between data and model, indicating the "stress" of the solution[2]. Such a stress-measure has been proposed by LEVELT (1970).

The psychology of grammar seems to be an area where it is only natural to expect hierarchical data structures. In phonology the idea of a hierarchical ordering of features is about as old as the concept of feature itself (JAKOBSON, 1941). Campbell's paper gives – again – psycholinguistic evidence. Syntax without phrase markers is unthinkable. Systematic introspection with respect to syntactic structure should reveal corresponding hierarchical relations. Levelt's relatedness scaling is a systematic method for collecting syntactic intuitions. It appears that such data can indeed be meaningfully related to hierarchical linguistic structure. In semantics, finally, dominance relations among semantic markers frequently occur, as has been argued by Katz and others. Judgments of meaning similarity may thus be expected to be at least partially ultrametric. MILLER's (1969) study supports this expectation.

In short, the psychology of grammar is a natural field of application for techniques of hierarchical data analysis. The following two articles exemplify such applications.

[2] Though defined differently this stress measure serves the same purpose as stress measures in multidimensional scaling (e.g. in KRUSKAL, 1964): to indicate goodness-of-fit of a solution.