New approaches to cluster analysis of typological indices

Michael Cysouw

1 Introduction

Ever since Greenberg's (1963) seminal paper on word order universals, there have been many approaches to investigate the relationship between typological parameters on the basis of large samples of the world's languages (cf. Cysouw 2005 for a survey). In the currently flourishing field of linguistic typology there is a strong consciousness about the need to include many languages in such comparisons, with the result that regularly hundreds of languages are included in typological studies. Notwithstanding such large sets of data, there has hardly been any progress in the quantitative analysis of the data collected. The only widely used method is the (intuitive) search for implicational universals, in the style as already used by Greenberg in 1963. There have been various proposals for different analytical methods, though unfortunately they have not caught on. In this paper, I will discuss one possible line of attack as first proposed by Altmann more than thirty years ago.

2 Hierarchical clustering

This novel approach to typological classification was first laid out in Altmann (1971), and further refined together with Lehfeldt in their joint publications (Altmann & Lehfeldt 1973, 1980). Altmann proposed to use the then newly developing clustering methods from biological phylogenetics for the analysis of typological variation. In the 1971 paper, Altmann investigated characteristics of the phonological system of Slavic languages. More precisely, for each language he constructed a 'phonological profile', a summary of phonological characteristics based on a feature-like analysis of Slavic sound systems, the details of which are not of importance here. The result is a set of twelve parameters, each describing some aspect of the phonological system of the Slavic languages. On the basis of these parameters, Altmann calculated the

difference for every pair of languages (using an Euclidean distance measure). He used these distances to construct a tree in which the most similar languages are put into branches of the same node of the tree. This tree is shown in Figure 1. For the construction of this tree, Altmann used the hierarchical clustering scheme proposed just a few years earlier by Johnson (1967). It is important to realize that with this tree Altmann did not attempt to reconstruct the historical relationships between the languages. The tree in Figure 1 is a typological classification. The tree-structure is only used to summarize relative similarities between the languages – similar languages being closer to each other in the tree structure. There is no claim as to the origin of these similarities.

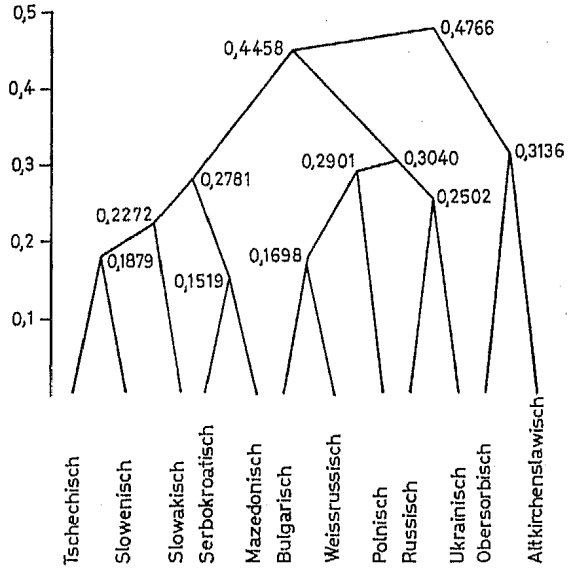

Figure 1: Hierarchical classification of Slavic phonological profiles (Altmann 1971: 19)

Building on Altmann's work, Lehfeldt (1972, cf. Altmann & Lehfeldt 1973: 39ff.; 1980: 282ff.) added an extra step of analysis to this approach. Once a classificatory tree is constructed, one might ask which parameters determine the subgrouping of languages. As an answer to this question, Lehfeldt (1972: 337) gives a long list of criteria on the basis of which the different branches of the tree can be distinguished. These criteria start at the

highest division of the tree and go down every node one by one. For every node a few parameters are listed that differentiate the groups of languages that split at this node.

The result of such an analysis is highly interesting for the field of linguistic typology. It provides an inductive method to classify languages into groups, and describes the linguistic characteristics of all these groups. Yet, to my knowledge, there have not been other typological investigation building on this work by Altmann and Lehfeldt. A similar approach, though apparently independently developed, has been applied by Sumie Ueda and Yoshiaki Itoh on word-order data collected by Tasaku Tsunoda (cf. Tsunoda et al. 1995, Ueda & Itoh 2002, Itoh & Ueda 2004). They also classify the languages in a tree-structure and investigate the primary division in the tree. They conclude that the coding of adpositions is the principle parameter explaining this division, and thus the main parameter to explain the word-order variation among the world's languages.

3 Problems

This approach opens up new possibilities to interpret typological data, but there are some problems that have to be addressed. The first problem with the tree as drawn by Altmann is that it looks like a tree as normally drawn in historical linguistics, with a root apparently indicating a prime division of groups. However, there is no inherent reason to pinpoint such a root on the basis of distance matrices. It is only possible to present a hierarchy of groups based on relative similarity. For that reason, the notion of an unrooted tree has been introduced in biological phylogenetics. An unrooted tree on the basis of the Slavic distances is shown in Figure 2.[1] There are nested groups indicating relative similarity, but there is no 'starting point' to read this tree.

The second problem is more problematic. A basic assumption of hierarchical clustering methods is that such a hierarchical clustering is actually possible. This is in many ways a problematic assumption, even more so when such methods are applied on often rather messy typological data. The result of this assumption is that the languages are forced into hierarchically organized groups, without acknowledging that the arguments for favoring one

1. To draw this tree, I used the Neighbour-Joining algorithm (Saitou & Nei 1987). Note that besides the absence of a root, there are also slight differences between this tree and the tree presented by Altmann as shown in Figure 1.

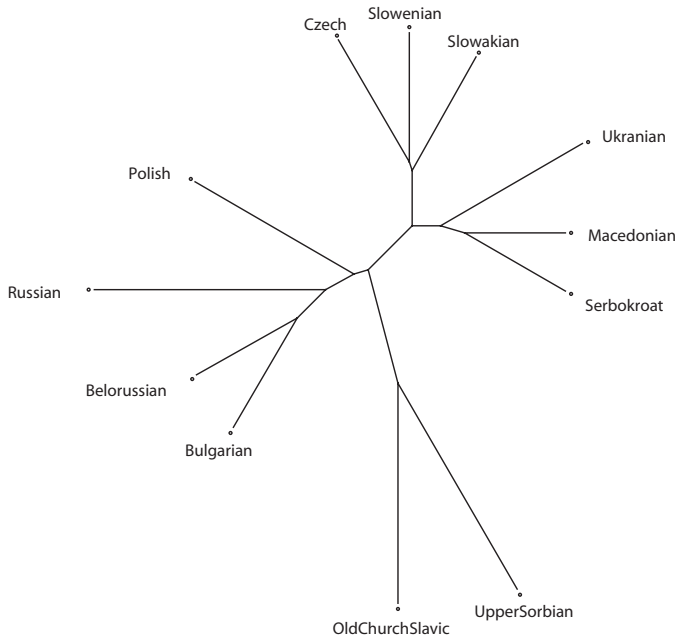

Figure 2: Unrooted tree of Slavic similarities, using the neighbour joining algorithm

particular grouping over another possibility are often not very strong. The impression of a clear hierarchical classification, as shown in Figure 1, might very well be deceptive (and I will argue that in the case of Slavic phonology it really is deceptive).

To exemplify this problem, consider the distances in Table 1 between the languages Czech, Slowakian, Serbo-Croatian and Macedonian, an excerpt from the data used by Altmann. The lowest distances (and thus the strongest similarity) are between the pairs Czech–Slowakian (0.1978) and Serbo-Croatian–Macedonian (0.1519), and indeed these languages are grouped together first, before they are both grouped together higher up in the tree (cf. Figure 1). However, as can be seen from the actual distances in Table 1, all values are rather close to each other. It might very well be some random effect that produced these small differences, which might imply that the lowest values are not significant.

Table 1: Normalised distances between selected Slavic languages

	Czech	Slowakian	Serbo-Croatian	Macedonian
Czech	0	0.1987	0.2366	0.2781
Slowakian	0.1987	0	0.2158	0.2310
Serbo-Croatian	0.2366	0.2158	0	0.1519
Macedonian	0.2781	0.2310	0.1519	0

4 Networks instead of trees

As an alternative, recent work in biological phylogenetics proposes various approaches that do not force binary trees, but allow for more network-like topologies when the data does not clearly suggest tree-like hierarchical clustering (cf. Huson & Bryant 2006 for a survey of various approaches to networks). I will here use the NeighborNet approach (Bryant & Moulton 2004) as implemented in SplitsTree4.[2] A NeighborNet interprets a distance matrix, and shows all possible groupings with branch-lengths proportional to the amount of evidence for such a grouping. For example, a simple NeighborNet is presented in Figure 3, based on the data from Table 1. This graph illustrates that there is both evidence for the grouping Macedonian–Serbo-Croatian versus Slowakian–Czech (a) and for the alternative grouping Macedonian–Slowakian versus Czech–Serbo-Croatian (b). However, the first grouping is clearly stronger, as can be seen from the longer lines of this side of the rectangle in the middle. From such a network no real clustering can be derived, though there is some indication that one of the groupings, namely (a), is indeed stronger than the others.

When this method is applied on the complete data from Altmann (1971), the result is a network as shown in Figure 4. The various sets of parallel lines in this graph indicate evidence for possible subgrouping. Upper Sorbian and Old Church Slavic rather clearly form a group. However, for the other languages there are no clear unique subgroupings. For example, Czech, Slowenian and Slowakian can be grouped with Serbo-Croatian, but also with Upper Sorbian and Old Church Slavonic, depending on which parallel lines are followed. There is no clear preference to be discerned for either of these groupings. Overall, there does not appear to be any clear grouping possible

2. SplitsTree4 is available at `http://www.splitstree.org`. The programm is described in Huson & Bryant (2006). An example of the use of NeighborNet on linguistic data can be found in Bryant et al. (2005).

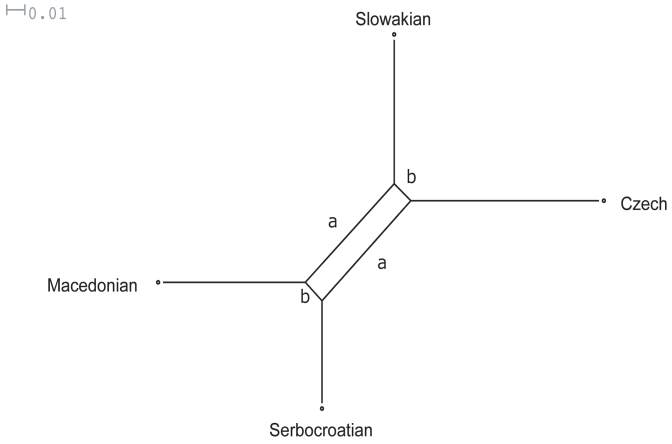

Figure 3: Neighbornet for four Slavic languages (data from Altmann 1971)

in this network. Concluding, the tree as drawn by Altmann is not a good summary of the distances between the languages. In that tree, the data are forced into a binary hierarchical subgrouping, but there is no real justification for this approach.

5 Using networks as a heuristic

Networks can not only be used to argue against a claimed classification. Conversely, they can also help to find exactly those divisions in the data that are worthwhile for further investigation. For example, consider some data from Altmann & Lehfeldt (1973: 40). The data describe morphological indices on the basis of text-counts, e.g. the number of inflections divided by the number of words. In total, there are ten such parameters distinguished.[3]

From this data-set a network is made, as shown in Figure 5. The network again does not show any clear clusters, except for the group of Jakut and

3. These parameters were first proposed by Greenberg (1990 [1954]), giving data for eight languages. A normalisation of Greenberg's parameters was proposed by Krupa (1965), adding also five more languages. These normalised parameters were used by Altmann & Lehfeldt (1973: 40), adding again data for eight more languages, but removing one of the languages added by Krupa because of lack of data.

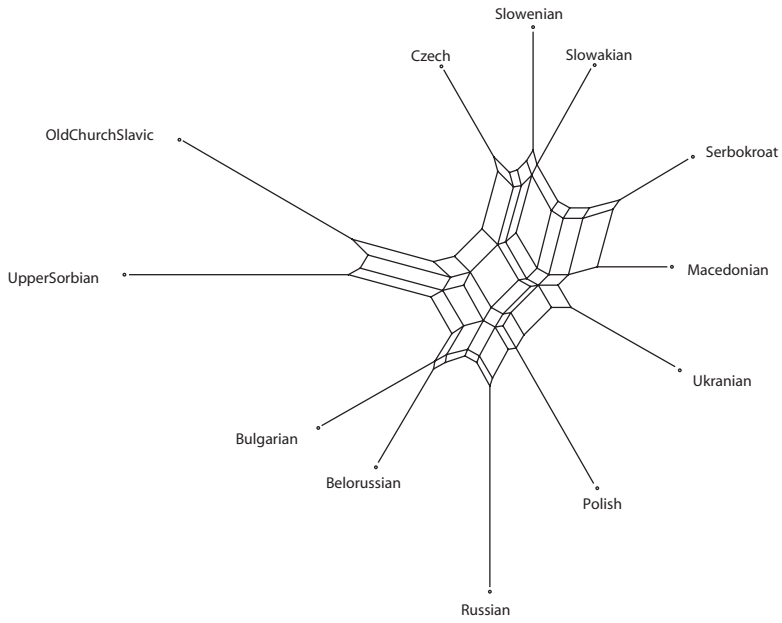

Figure 4: NeighborNet of Slavic phonological profiles (based on the data from Altmann 1971: 14, Table 4)

written Turkish (which is an interesting observation in itself, but this will not be further pursued here).

Yet, note that the various non-European languages are placed rather outside of the network (Eskimo, Vietnamese, Swahili, Jakut, written Turkish). The long lines separating these languages from the other languages indicate that they are all individually very dissimilar from the rest. Nothing can really be deduced from such long individual branches. When they are subsequently removed, the result is a much more interesting network, as shown in Figure 6. This network shows a clear division in the middle, separating the modern languages from the languages that are only attested through historical records.[4] On the basis of this network, a hypothesis can now be formulated proposing that there is a significant morphological difference between the modern and the old languages, giving rise to a typological division of these languages into two groups. There are also some indication of possible smaller groups (e.g.

4. For reasons of clarity of the presentation, Hethitic has also been removed from this network. Hethitic appears roughly in the middle between the two apparent groups.

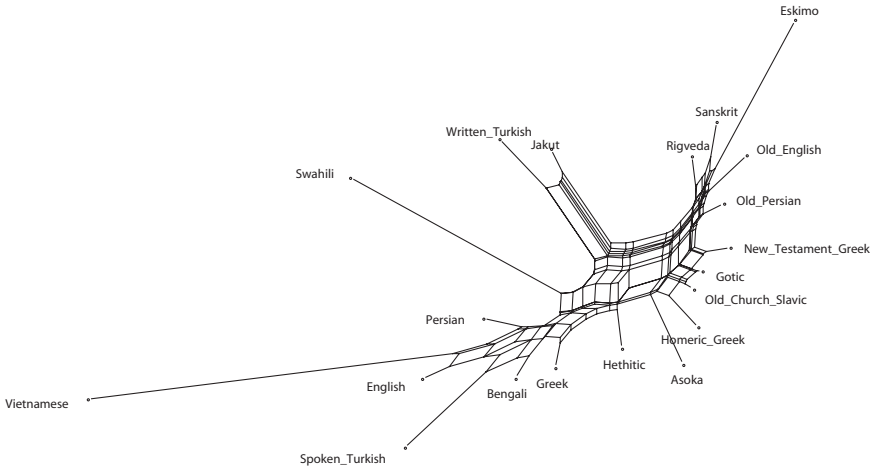

Figure 5: NeighborNet of the similarities between typological profiles (based on data from Altmann & Lehfeldt 1973: 40)

Old Persian, Old English, Sanskrit and Rigveda), but they will not further be investigated here.

6 Characterising subclusters

On the basis of the hypothesized typological division, it is now possible to investigate the reasons why the languages in these two groups are different. Which parameters distinguish these two apparent groups of old and new languages? Instead of looking for potential crucial parameter settings (as attempted by Lehfeldt), I have compared the two groups of languages for all parameters using *t*-tests. It turns out that the two groups of languages are significantly different on almost all parameters – the only exception being Compounding.

These differences can be clearly distinguished in the box-plot as shown in Figure 7. In this box-plot, the value-variation for all parameters is shown, the old languages being shown in white and the new languages in grey (Compounding is omitted). The whiskers of each box indicate the minimum and the maximum of variation, ignoring outliers, which are given as small open circles. In all cases, the whiskers are non-overlapping, showing that the variation between the two groups on all parameters is clearly separated. The

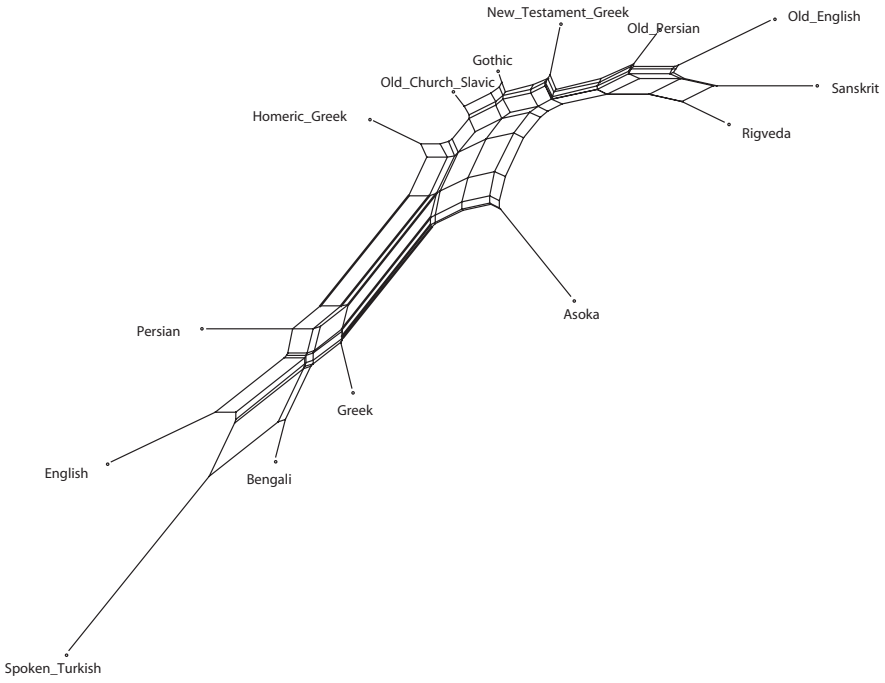

Figure 6: NeighborNet from same data as in Figure 2, with outliers removed

important conclusion from this example is that when there are clear groups to be distinguished typologically, then there are probably not just a few parameters that make the difference. Substantially different groups will be reflected in many parameters, or even in almost all.

Let me spend just a few words on the meaning of the parameters in the current example. The modern languages are generally higher on Analyticism, Agglutination and Isolation. Analyticism is defined as the number of words divided by the number of morphemes. The higher values for the modern languages indicate that the modern languages have relatively less morphemes per word. Agglutination is defined as the fraction of morpheme boundaries that show no alternation or only phonologically regulated alternation. The higher values for the modern languages indicate less morphologically governed allomorphy and suppletion.[5] Isolation is defined as the fraction of grammati-

5. The influence of modern Turkish (a prototypical agglutinative language) on this result is limited. Turkish is an outlier in the boxplot, indicated by the little dot at 0.67.

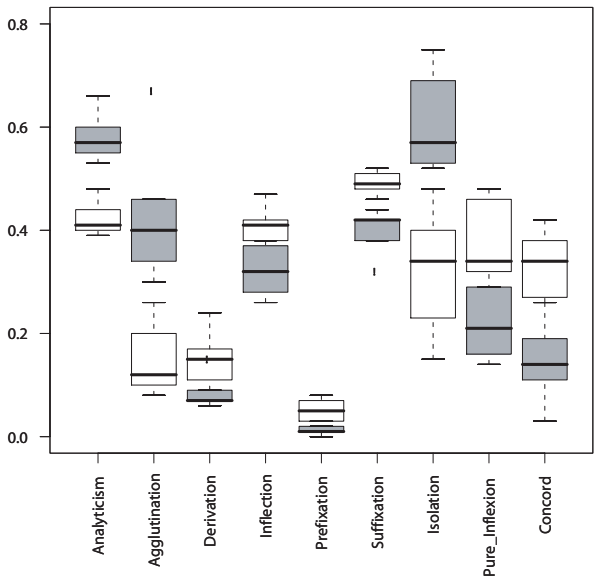

Figure 7: Boxplot comparing typological characteristics between the extinct languages (white) to the contemporary languages (grey)

cal relations that is marked by overt morphology. The higher values for the modern languages indicate that they have more zero-marked grammatical relations (though often marked by word order). In contrast, for all parameters measuring the use of various kinds of morphological material (derivation, inflection, prefixation, suffixation, concord) the modern languages are significantly lower than the old languages. This division of parameters into two groups makes good sense, and describes roughly the traditional morphological types flectional (old) against isolating/agglutinating (modern).

7 Relation between parameters

In the approach as described until now, the languages are grouped together first, and only then, in a second step, the linguistic parameters that characterize these groups are investigated. This procedure can of course be turned upside down by first grouping the parameters and then grouping the languages. This would be closer to the practice of linguistic typology where first language types are identified and then languages are classified as belonging to a

particular type. However, depending on the data this is not always a profitable approach. For example, I will argue that the morphological data discussed previously do not give good results under the reversed method.

Following Altmann's approach, I have made a hierarchical clustering of the parameters, as shown in Figure 8.[6]

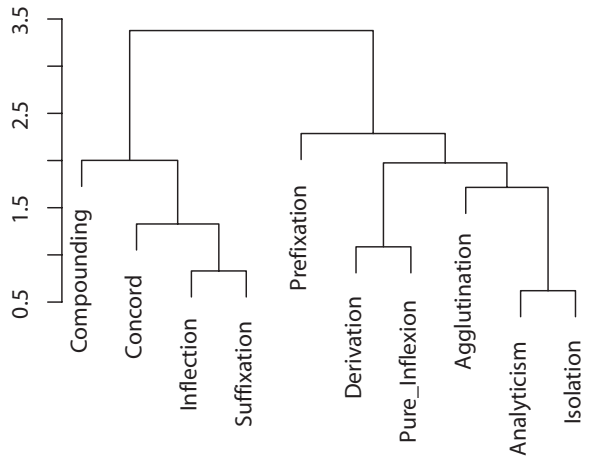

Figure 8: Hierarchical clustering of the Greenbergian morphological parameters

The parameters appear to fall in two groups as indicated by the highest level branch of the classification. Unfortunately, these two groups do not coincide with the results from the previous section (in which it was argued that the parameters Analyticism, Agglutination and Isolation form a group, to be found as a subcluster at the right side of Figure 8). However, as I have argued in Section 3, a forced-binary classification can create an incorrect impression with messy data. Indeed, the NeighborNet in Figure 9 shows that the classification from Figure 8 might be a good possibility if one is forced to

6. To compute distances between parameters, I have normalised the values for each parameter to fill out the whole range between 0 and 1. For every value v from parameter P the normalised value is $v - min(P)/max(P) - min(P)$. This normalisation was necessary because some parameters (e.g. Compounding) had high values for all languages, but the values for most other parameters were more spread out. The effect is that Compounding would end up as being very different from all other parameters. What is needed here is not the absolute distance between the values of two parameters, but a relative distance (i.e. including the internal variation within one parameter). The classification as shown in Figure 8 was made using the routine *hclust* from the statistical package *R*.

choose a binary classification, but there are very many alternative possibilities to group the parameters. The strongest split in Figure 9 separates the parameters Prefixation, Analyticism, Isolation and Agglutination from the rest. This is almost the same set of parameters as identified in the previous section (pace Prefixation). However, overall there is not really strong evidence in this network to separate out groups of parameters.

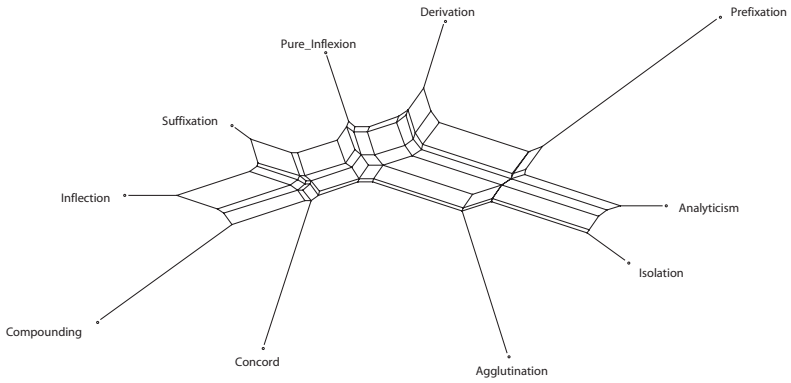

Figure 9: NeigborNet of the Greenbergian morphological parameters

The parameter-grouping as found in the previous section can be recovered when yet another visualization tool is used: multidimensional scaling, (MDS, cf. Cysouw 2001 and Croft 2004 for the application of MDS on typological data). In multidimensional scaling the real distances between the parameters are tweaked slightly until they can be fit into a lower-dimensional graph (preferably one- or two-dimensional). A two-dimensional MDS of the morphological parameters is shown in Figure 10.[7] This visualization of the data coincides with the results from the previous section. The three parameters Analyticism, Agglutination and Isolation are grouped closely together to the lower left. Almost all other parameters are found together in the top of the MDS, except for Compounding, which is found separately from all other parameters in the lower right.

7. The MDS in Figure 10 was made by using the routine *cmdscale* from the statistical package *R* on the normalised parameter values as described in footnote 6.

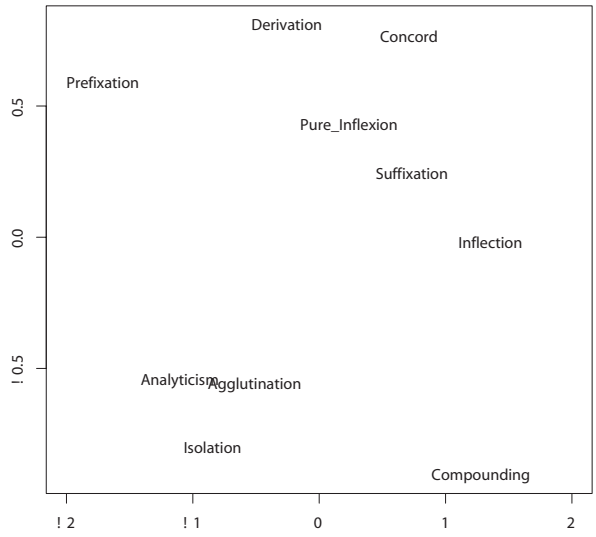

Figure 10: Multidimensional scaling of the Greenbergian morphological parameters

8 Conclusion

Methods for clustering data are extremely useful to investigate typological indices. This idea has already been around for a few decades now, but not much use has been made of such methods in typological investigations. However, until recently it was rather difficult for typologists (which are often not strongly mathematically oriented) to use such methods, as they were not very accessibly implemented. In recent years, various more easy to use implementations have become available. Unfortunately, it is still necessary to switch between various software packages, with all encoding problems that are involved in such transitions, to use the methods as described in this paper. Real progress in the quantitative analysis of typological data will probably only be made if a ready-made software package for typologists is compiled, in which all methods that are useful for typology are combined in an easy to use interface.

References

Altmann, Gabriel
1971 "Die phonologische Profilähnlichkeit. Ein Beitrag zur Typologie phonologischer Systeme der slawischen Sprachen". In: *Phonetica*, 24; 9–22.

Altmann, Gabriel; Lehfeldt, Werner
1973 *Allgemeine Sprachtypologie*. München: Fink.
1980 *Einführung in die Quantitative Phonologie*. Bochum: Brockmeyer.

Bryant, David; Filmon, Flavia; Gray, Russell D.
2005 "Untangling our past: Languages, trees, splits and networks". In: Mace, Ruth; Holden, Clare J.; Shennan, Stephan (Eds.), *The Evolution of Cultural Diversity: A Phylogenetic Approach*. London: UCL, 67–84.

Bryant, David; Moulton, Vincent
2004 "Neighbor-Net: An agglomerative method for the construction of phylogenetic networks". In: *Molecular Biology and Evolution*, 21(2); 255–265.

Croft, William; Poole, Keith T.
2004 "Inferring universals from grammatical variation: multidimensional scaling for typological analysis". [*Unpublished Manuscript*].

Cysouw, Michael
2001 "[Rev.:] Martin Haspelmath (1997): Indefinite Pronouns". In: *Studies in Language*, 37(3); 99–114.
2005 "Quantitative methods in typology". In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Raimund (Eds.), *Quantitative Linguistics: An International Handbook*. Berlin: Mouton de Gruyter, 554–578.

Greenberg, Joseph H.
1963 "Some universals of grammar with particular reference to the order of meaningful elements". In: Greenberg, Joseph H. (Eds.), *Universals of Language*. Cambridge, Mass.: MIT Press, 73–113.
1954 "A quantitative approach to the morphological typology of language". In: Denning, Keith; Kemmer, Suzanne (Eds.), *On Language: Selected Writings of Joseph H. Greenberg*. Repr. 1990. Stanford, CA: Stanford University Press, 3–25.

Huson, Daniel H.; Bryant, David
2006 "Application of phylogenetic networks in evolutionary studies". In: *Molecular Biology and Evolution*, 23(2); 254–267.

Itoh, Yoshiaki; Ueda, Sumie
2004 "The Ising model for changes in word ordering rules in natural languages". In: *Physica D*, 198(4); 333–339.

Johnson, Stephen C.
1967 "Hierarchical clustering schemes". In: *Psychometrika*, 32(3); 241–254.
Krupa, Viktor
1965 "On quantification of typology". In: *Linguistics*, 12; 31–36.
Lehfeldt, Werner
1972 "Phonologische Typologie der slavischen Sprachen". In: *Die Welt der Slaven*, 17; 318–340.
Saitou, Naruya; Nei, Masatoshi
1987 "The neighbour-joining method: A new method for reconstructing phylogenetic trees". In: *Molecular Biology and Evolution*, 4(4); 406–425.
Tsunoda, Tasaku; Ueda, Sumie; Itoh, Yoshiaki
1995 "Adpositions in word-order typology". In: *Linguistics*, 33; 741–761.
Ueda, Sumie; Itoh, Yoshiaki
2002 "Classification of natural languages by word ordering rule". In: Opitz, Otto; Schwaiger, Manfred (Eds.), *Explanatory Data Analysis in Empirical Research*. Berlin: Springer, 180–187.