A quantitative approach to Dutch verb cluster variation  
Jeroen van Craenenbroeck (KU Leuven)  
jeroen.vancraenenbroeck@kuleuven.be

SUMMARY This paper combines quantitative-statistical and formal-theoretical approaches to language variation. I provide a quantitative analysis of verb cluster variation in 267 dialects of Dutch and map the results of that analysis against grammatical microparameters extracted from the theoretical literature. Based on this new methodology, I argue that variation in verb cluster ordering in Dutch dialects can be largely reduced to three grammatical microparameters.

THE DATA: VARIATION IN DUTCH VERB CLUSTERS Of the 6 theoretically possible orderings for the three-verb cluster in (1), 4 are attested in Dutch dialects (Barbiers et al., 2008):

(1) a. Ik vind dat iedereen moet kunnen zwemmen.  
I find that everyone must can swim  
‘I think everyone should be able to swim.’

b. Ik vind dat iedereen zwemmen kunnen.  
c. Ik vind dat iedereen zwemmen kunnen moet.  
d. Ik vind dat iedereen kunnen zwemmen moeten.  
e. *Ik vind dat iedereen kunnen zwemmen moeten.  
f. *Ik vind dat iedereen moeten kunnen zwemmen.

However, not every Dutch dialect allows every one of the orders in (1a-d): some allow one, some 2, some 3, and some 4. In total, there are 12 different combinations of orders attested in Dutch dialects, i.e. there are 12 different dialect types for this single verb cluster. When we do the same analysis for the 31 verb cluster orders that were investigated in Barbiers et al. (2008), we find a staggering 137 different types of dialects.

RESEARCH QUESTIONS This high degree of interdialectal variation raises fundamental questions for parameter theory: to what extent is this variation determined by grammatical microparameters, and how do we go about identifying those parameters?

THE METHODOLOGY The analysis starts from a 31×267 data table with cluster orders as rows and dialect locations as columns. Cells contain “yes” if that cluster order occurs in that dialect, and “no” if it doesn’t. To this table I apply a Multiple Correspondence Analysis (Husson et al., 2011), which involves two steps. First, the raw data table is transformed into a 31×31 distance matrix, whereby each cluster order is compared to each other cluster order and the degree of difference between them is given a score between 0 (= identical geographical distribution) and 1 (= complementary geographical distribution). Second, the 31×31 distance matrix is reduced to a two- or three-dimensional one for easy visualization and interpretation. A two-dimensional representation of the verb cluster data under investigation here is given in figure 1. In this graph, each of the 31 cluster orders is situated on a two-dimensional plane. When two cluster orders are close together this means that they have a highly similar geographical spread, while when two orders are far apart they typically do not occur in the same dialect locations.

Figure 1: Two-dimensional plot of 31 verb cluster orders
The question now is to what extent this variation can be accounted for by grammatical factors. In order to investigate this I extracted possible grammatical parameters from the theoretical literature. For example, in Barbiers (2005)'s analysis, the base order is strictly head-initial, all movements are VP-intrapositions, they are feature-driven and can pied-pipe VPs other than the one undergoing checking. This means we can now categorize verb clusters according to whether (a) they can be base-generated, (b) they involve movement, (c) they involve pied-piping, and (d) they involve a feature-checking violation. A small sample is given in Table 1.

<table>
<thead>
<tr>
<th>feature-checking</th>
<th>base-generation</th>
<th>movement</th>
<th>pied-piping</th>
<th>violation</th>
</tr>
</thead>
<tbody>
<tr>
<td>auxiliary-participle</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>participle-auxiliary</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>infinitive-modal</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 1: Partial classification of cluster orders based on Barbiers (2005)

I have added 70 such linguistic variables to the analysis, representing the analyses of among others Barbiers (2005), Barbiers and Bennis (2010), Abels (2011), Haegeman and Riemsdijk (1986), Bader (2012), and Schmid and Vogel (2004). We can then investigate which of those variables provides the best match for the empirical variation depicted in figure 1. For example, is there a linguistic parameter such that all the negative values on the x-axis in figure 1 correspond to one setting of that parameter, and all the positive values to the other setting? In order to measure the degree of correspondence between the data and the theory, we calculate the squared correlation ratio ($\eta^2$), which is a measure for the proportion of variance along one of the dimensions of figure 1 that is explained by a specific linguistic variable.

**THE RESULTS: THREE GRAMMATICAL MICROPARAMETERS** Reducing the $31 \times 31$ distance matrix to three (rather than two, as in figure 1) dimensions accounts for 78.46% of the variation found in the verb cluster data. I take this to mean that there are three grammatical microparameters at play in this data set, and that these three parameters together are responsible for roughly 80% of the variation. In order to identify those parameters we need to map each of those three dimensions against the 70 variables extracted from the theoretical literature and see which ones match up best. The first dimension turns out to be highly correlated with the position of participles and infinitives vs. their selecting verbs, i.e. auxiliaries and modals ($\eta^2=0.61$): it sets apart dialects where the modal precedes the infinitive and the participle precedes the auxiliary from dialects where at least one of these conditions is not met. The second dimension sets apart clusters that end in a descending order (21, 132, 321, 231) from clusters that end in an ascending order (12, 123, 312, 213) ($\eta^2=0.38$). The third dimension separates strictly head-final orders (21 and 321) from all the others ($\eta^2=0.68$).

**BEYOND THE NUMBERS: A PARAMETRIC ANALYSIS** Based on these three parameters we can construct a parametric account of verb cluster ordering in Dutch, whereby we start out from a head-final base order, which can be changed by movement or not (dimension 3: ±MOVEMENT); if it is, this can involve multiple movement steps in the case of three-verb clusters (123, 312, 213) or not (321, 132, 231) (dimension 2: ECONOMYOFMOVEMENT); and this is combined with a head parameter regulating the position of participles and infinitives (dimension 1).

**FURTHER EXTENSIONS: HEADEDNESS** This methodology can shed new light on old verb cluster puzzles such as headedness: we can directly compare head-initial, head-final, and ‘mixed’ analyses and see to what extent the patterns they predict also show up in the data.