

**Descriptive, methodological and theoretical issues in syntactic variation research
A multivariate corpus analysis of word order variation in Dutch clause final verb
clusters**

*Gert De Sutter, Dirk Speelman & Dirk Geeraerts
RU Quantitative Lexicology and Variational Linguistics
Department of Linguistics - University of Leuven
Blijde-Inkomststraat 21
3000 Leuven, Belgium*

{ gert.desutter,dirk.speelman,dirk.geeraerts }@arts.kuleuven.ac.be

In recent years, there has been enormous progress in the field of syntactic variation research, yielding a better insight into the variables that determine the choice between syntactic alternatives and a better understanding of human sentence processing (e.g. Hawkins 1994; Arnold et al. 2000; Clifton & Duffy 2001; Wasow 2002; Gries 2003). Nevertheless, some types of syntactic alternation remain a tough nut to crack, in that well-known determiners such as syntactic complexity, semantic-pragmatic value and ambiguity avoidance, which can be explained in terms of communicative or on-line processing needs (Arnold et al. 2000), do not seem to account for the variation.

In the present study, we focus on such a case of syntactic variation that is well-known for its inextricable tangle of interacting forces, viz. **word order variability in Dutch clause final verb clusters**. Consider the following pair of examples:

[...] dat	hij	de afwas	gedaan	heeft / heeft	gedaan.
[...] that-COMP	he-SUBJ	the dishes-OBJ	done-PART	has-AUX/ has-AUX	done-PART.
[...] that he has done the dishes.					

Both word order variants *gedaan heeft* (participle-first word order) and *heeft gedaan* (participle-final word order) are equally grammatical, and do not seem to have any obvious semantic-pragmatic consequences. This type of word order variation has been studied intensively since the 1950's. Many investigations made clear that this type of word order variation is not a case of free variation, as numerous language-structural as well as contextual variables influence the choice of word order (see De Sutter 2005 for an overview). Nevertheless, several questions remain unanswered: (i) how do the already introduced variables relate to each other in their effect on the choice of word order? More particularly, what is the relative effect of each variable, i.e. the effect of a variable, given the effect of all other variables? Are some of the variables redundant, given the effect of one or more other variables? (ii) What is the global effect and the predictive power of these variables? Formulated somewhat differently, how much of the observed observation can be explained by the variables and to what extent can word order be predicted on the basis of the variables? This lack of clarity with respect to the mutual relationships of the variables and the simultaneous effect of all variables together has prevented the identification of the set of variables that determines the choice of word order and has obstructed a linguistic interpretation of the word order variation at hand (why do both syntactic variants coexist; what is the functional difference between both syntactic variants?).

Building on a large-scale representative corpus of contemporary Dutch (CONDIV-corpus; 45 million tokens), this study will try to answer these questions by **statistically modelling the word order variation** in Dutch clause final verb clusters **on the basis of 10 language-internal** (prosodic, morphosyntactic, semantic,

discursive) **variables**, while controlling for several language-external variables. In order to achieve that goal, logistic regression analysis is used (cf. Agresti 1996).

The results of the logistic regression analysis show that:

1. **8 out of 10 predictor variables significantly contribute** to the choice of word order in Dutch clause final verb clusters. More particularly, all morphosyntactic, semantic and discursive variables that were included in the model are statistically significant, whereas both prosodic variables do not significantly contribute to the explanation of the word order.
2. **The semantic variable is the most important variable** in the model (odds ratio = 18.30).
3. The set of 8 predictor variables significantly **explains the bulk of the variation** in our data set: the deviance reduction of the informed model (i.e. the model containing 8 predictor variables), compared to the dummy model (containing no predictor variables), equals 706.1 (= 3031.8 – 2325.7). This reduction is statistically significant ($p < .05$).
4. The model is able to **predict the choice of word order satisfactorily**: the concordance measure (c), which gives an index for the correlation between the predicted probabilities and the observed values of the response variable, equals 0.803 (after 100 bootstrap repetitions), so that one can conclude that “A model having c greater than roughly .8 has some utility in predicting the responses of individual subjects” (Harrell 2001: 247).

For the first time in over 50 years of empirical research into word order variation in Dutch clause final verb clusters, we have been able to detect and weight the set of variables that governs the choice of word order in Dutch clause final verb clusters. One of the most striking findings emerging from this investigation is undoubtedly the multivariate nature of the word order variation at hand. Since this is not an incidental pattern in syntactic variation research (cf. the analyses conducted by Grondelaers 2000, Gries 2003, Wulff 2003), we will deal with two obvious implications of this multivariate nature on a more general methodological and theoretical level:

1. Since there is no predetermined set of factors to include in syntactic variation research, researchers have been selecting and testing the influence of **randomly selected variables**. As a consequence, a coherent view on the (generally valid) determining forces of syntactic variation is lacking. We will argue that researchers should select variables more systematically, more particularly by taking into account structurally diverse types of variables (i.e. phonological, prosodic, morphosyntactic, semantic, discursive and lectal factors) when statistically modelling syntactic variation, either by including them in the model or by controlling for the effect of the non-included factors. Given the multivariate nature of syntactic variation phenomena, one should ask the question to what extent the traditional conception of a theory of grammar should be modified in order to account for this multivariate reality, in which different types of functional features (structural/semantic, textual, lectal) co-determine the presence of linguistic features. We will argue that such a **multivariate grammar model** can be developed in the context of usage-based approaches to language and language variation (cf. Barlow & Kemmer 2000).

References

- Agresti, A. (1996). *An introduction to categorical data analysis*. New York et al.: Wiley.
- Arnold, J.E., T. Wasow, A. Losongco & R. Ginstrom (2000). "Heaviness vs. newness: the effects of structural complexity and discourse status on constituent ordering." *Language* 76, 28-55.
- Barlow, M. & S. Kemmer (2000). *Usage-based models of language*. Stanford: CSLI.
- Clifton, C. Jr. & S.A. Duffy (2001). "Sentence and text comprehension: Roles of linguistic structure". *Annual Review of Psychology* 52, 167-196.
- Gries, Stefan Th. (2003). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London / New York: Continuum Press.
- Grondelaers, S. (2000). *De distributie van niet-anaforsch er buiten de eerste zinsplaats. Sociolinguïstische, functionele en psycholinguïstische aspecten van er's status als presentatief signaal*. PhD K.U. Leuven.
- Hawkins, J.A. (1994). *A performance theory of order and constituency*. Cambridge: CUP.
- Sutter, G. De (2005). *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen*. PhD K.U.Leuven.
- Wasow, T. (2002). *Postverbal behavior*. Stanford: CSLI.
- Wulff, S. (2003). "A multifactorial corpus analysis of adjective order in English". *International Journal of Corpus Linguistics* 8, 245-82.

Appendix: Logistic regression analysis

Predictor variables (the formal notation of the variable, together with the values it can take, are given between brackets; in case of a categorical variable, the reference value is underlined): *type of auxiliary* (aux; values: copula vs. time vs. passive *zijn* vs. *worden* vs. unclassifiable), *frequency of the participle* (freq.part; values: numerical), *morphological structure of the participle* (morph.part; values: non-separable vs. separable), *length of the middle piece* (mid.piece; values: 0-2 words vs. 3-5 words vs. 6-8 words vs. 12-14 words vs. >14 words), *syntactic persistence* (syn.persistence: none vs. green word order versus red word order), *informationality of the last preverbal word* (informat; values: low vs. intermediate vs. high), *inherence of the last preverbal word* (inherence; fixed expression vs. non-fixed expression), *postverbal constituent* (pvc; values: complement of V vs. complement of N vs. adjunct/no pvc), *distance between the participial word accent and the last preverbal word accent* (unacc.syll.pre; values: 0&1 syllable(s), 2&3 syllable(s), >3 syllable(s)), and, finally, *distance between the participial word accent and the first postverbal word accent* (unacc.syll.post; values: 0&1 syllable(s), 2&3 syllable(s), >3 syllable(s)).

Model fit: $p < .05$

V.I.F. < 10 (for all included variables)

Predictor variable	β	ASE	z-value	Pr (> z)	O.R.
constant	-3.689e+00	4.055e-01	-9.096	< 2e-16 ***	
aux: time	2.907e+00	2.337e-01	12.437	< 2e-16 ***	18.30
aux: passive <i>zijn</i>	2.057e+00	2.535e-01	8.113	4.94e-16 ***	7.82
aux: <i>worden</i>	2.462e+00	2.296e-01	10.722	< 2e-16 ***	11.73
aux: unclassifiable	1.468e+00	3.389e-01	4.332	1.48e-05 ***	4.34
freq.part	2.441e-06	7.738e-07	3.15	0.0016 **	
morph.part : separable	1.352e+00	1.821e-01	7.426	1.12e-13 ***	3.87
mid.piece: 3-5 words	7.083e-01	1.459e-01	4.856	1.20e-06 ***	2.03
mid.piece: 6-8 words	8.288e-01	1.613e-01	5.139	2.76e-07 ***	2.29
mid.piece: 9-11 words	8.270e-01	2.009e-01	4.116	3.85e-05 ***	2.29
mid.piece: 12-14 words	9.446e-01	3.001e-01	3.147	0.001649 **	2.57
mid.piece: >14 words	6.832e-01	3.653e-01	1.871	0.061407 .	1.98
syn.persistence: none	5.404e-01	1.339e-01	4.035	5.47e-05 ***	1.72
syn.persistence: red	1.190e+00	1.363e-01	8.730	< 2e-16 ***	3.28
informat.: intermediate	3.434e-01	2.051e-01	1.674	0.094125 .	1.41
informat.: high	6.625e-01	1.991e-01	3.327	0.000878 ***	1.94
inherence: fixed expr.	8.163e-01	1.850e-01	4.412	1.02e-05 ***	2.26
pvc: complement of V	-7.423e-01	1.581e-01	-4.694	2.68e-06 ***	0.47
pvc: complement of N	1.947e-01	2.593e-01	0.751	0.452740	
#unacc.syll.pre: 0&1	-1.543e-01	1.874e-01	-0.823	0.410371	
#unacc.syll.pre: 2&3	1.060e-01	1.810e-01	0.586	0.558147	
#unacc.syll.post: 0&1	4.485e-03	2.231e-01	0.020	0.983956	
#unacc.syll.post: 2&3	-2.679e-02	1.863e-01	-0.144	0.885671	