



# A multivariate approach to the description and modeling of word order variation

Gert De Sutter,  
Dirk Speelman & Dirk Geeraerts



University of Leuven  
RU Quantitative Lexicology and Variational Linguistics

# General introduction

- Increased interest in grammatical variation research in recent years:
  - dative alternation
  - extraposition
  - relative order of nominal/prepositional constituents

# General introduction

- Increased interest in grammatical variation research in recent years:
  - dative alternation
  - extraposition
  - relative order of nominal/prepositional constituents
- Yielding a better insight into some motivating principles of grammatical optionality:
  - avoid ambiguity
  - mark pragmatic value
  - postpone complex constituents



# General introduction

- Nevertheless, some types of grammatical alternation remain a tough nut to crack

# General introduction

- Nevertheless, some types of grammatical alternation remain a tough nut to crack:
  - word order variation in Dutch clause final verb clusters:
    - [...] dat Jan dat boek heeft<sub>AUX</sub> gestolen<sub>PART</sub>.
    - [...] dat Jan dat boek gestolen<sub>PART</sub> heeft<sub>AUX</sub>.

# General introduction

- Nevertheless, some types of grammatical alternation remain a tough nut to crack:
  - word order variation in Dutch clause final verb clusters:
    - [...] dat Jan dat boek heeft<sub>AUX</sub> gestolen<sub>PART</sub>.
    - [...] dat Jan dat boek gestolen<sub>PART</sub> heeft<sub>AUX</sub>.

- one of the alternating elements is semantically empty
  - 
  -

# General introduction

- Nevertheless, some topics remain a tough nut to crack
  - word order variation in German
    - [...] dat Jan dat boek
    - [...] dat Jan dat be
  - one of the alternating elements is semantically empty
  - interconstituent variation vs. intraconstituent variation
- Mostly **interconstituent variation**:
  - The waiter brought the wine we had ordered to the table.
  - The waiter brought to the table the wine we had ordered.
- What about **intraconstituent variation**?
  - Less well-studied
  - More complex type of variation
  - Other motivating principles

# General introduction

- Nevertheless, some topics remain a tough nut to crack
    - word order variation in German
      - [...] dat Jan dat boek
      - [...] dat Jan dat bo
    - different methods
    - different empirical fundamentals
    - unclear data selection
    - no / restricted use of statistical techniques
- gestoien PART THEE/AUX.
- one of the alternating elements is semantically empty
  - inter-constituent variation vs. intraconstituent variation
  - methodological and analytical problems

# Purpose

- Model grammatical variation in Dutch verb clusters
- Contribute to the recent trend towards a fully-fledged theoretical and methodological framework for usage-based grammatical variation research

# Purpose

- Model grammatical variation in Dutch verb clusters:
  - Identify and weight the underlying mechanisms on the basis of which the grammatical alternation can be described, explained and predicted
- Contribute to the recent trend towards a fully-fledged theoretical and methodological framework for usage-based grammatical variation research

# Purpose

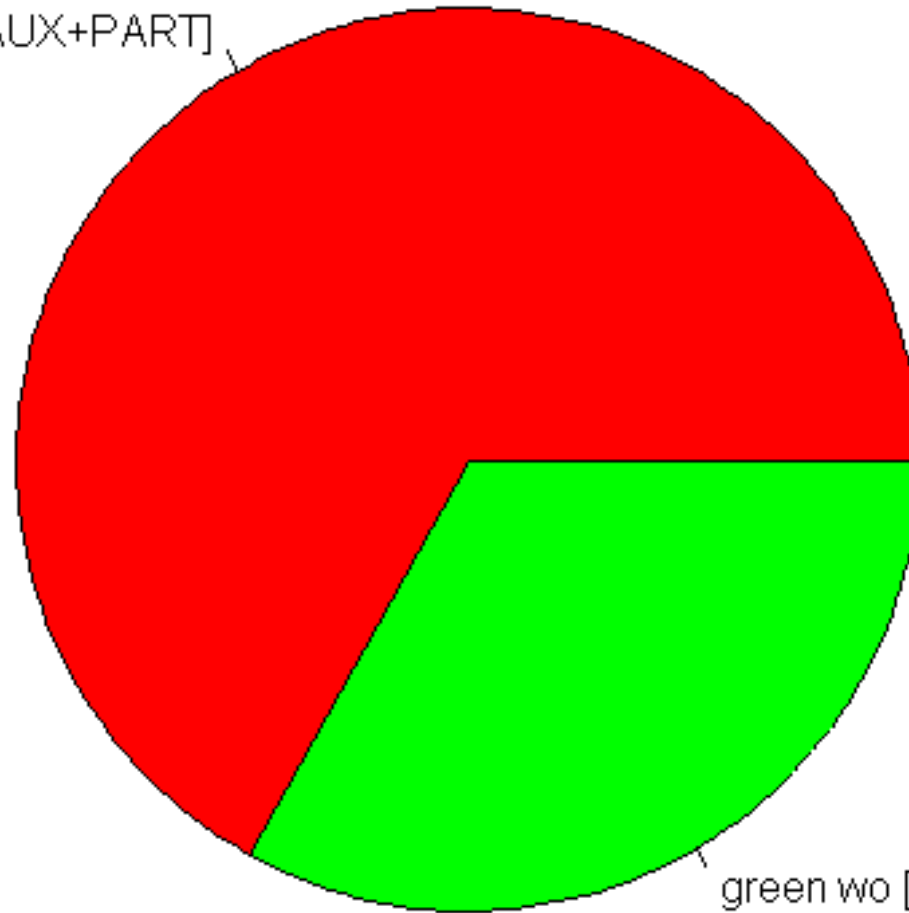
- Model grammatical variation in Dutch verb clusters:
  - in actual language use
  - from a synchronic perspective
  - with considerable attention for methodological and analytical prerequisites of thorough empirical research
  - using multivariate statistical techniques
- Contribute to the recent trend towards a fully-fledged theoretical and methodological framework for usage-based grammatical variation research

# Design of the study

- Corpora: ConDiv (subcomponent: ds\_supra)
- Data selection:
  - bipartite verb clusters, consisting of a participle and the auxiliary verb *zijn*, *hebben* or *worden*
  - complement clause introduced by *dat*
- Data extraction and processing: Abundantia Verborum
  - n = 2390
- Statistical analyses: R 2.0.1 (2003)

# Design of the study

red wo [AUX+PART]



green wo [PART+AUX]

# List of explanatory variables

1. Morphology of the participle
2. Frequency of the participle
3. Semantic-categorial status of the participle
4. Informationality of the preceding word
5. Inherence of the preceding constituent
6. Length of the middle piece
7. Syntactic function of the extraposed constituent
8. Distance between preceding word accent and participial word accent (i.t.o. syllables)
9. Distance between following word accent and participial word accent (i.t.o. syllables)
10. Syntactic persistence

# Multivariate statistical analysis

- Statistical modeling of the variation:
  - Do all exp. variables have a significant effect on the choice of word order?
  - What is the relative impact of each exp. variable?
  - How do the different exp. variables relate to each other?
  - What is the collective effect of all exp. variables on the choice of word order?
  - What is the predictive power of the model?

# Multivariate statistical analysis

- Logistic regression analysis:

$$\log \left[ \frac{p(Y=1)}{p(Y=0)} \right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots$$

- Modeling odds in favour of **red word order [AUX+PART]**

<b>Predictor</b>	<b><math>\beta</math></b>	<b>ASE</b>	<b>z-value</b>	<b>Pr(&gt; z )</b>	<b>O.R.</b>
intercept	-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:pass_zijn	2.057e+00	2.535e-01	8.113	4.94e-16 ***	7.82
Aux:worden	2.462e+00	2.296e-01	10.722	< 2e-16 ***	11.73
Aux:unclass	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:sep	1.352e+00	1.821e-01	7.426	1.12e-13 ***	3.87
mid_p: 3-5	7.083e-01	1.459e-01	4.856	1.20e-06 ***	2.03
mid_p: 6-8	8.288e-01	1.613e-01	5.139	2.76e-07 ***	2.29
mid_p: 9-11	8.270e-01	2.009e-01	4.116	3.85e-05 ***	2.29
mid_p: 12-14	9.446e-01	3.001e-01	3.147	0.001649 **	2.57
mid_p: >14	6.832e-01	3.653e-01	1.871	0.061407 .	1.98
synt_pers: none	5.404e-01	1.339e-01	4.035	5.47e-05 ***	1.72
pre-eindgroep: red	1.190e+00	1.363e-01	8.730	< 2e-16 ***	3.28
informat.: intermed.	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
bigram: sign	8.163e-01	1.850e-01	4.412	1.02e-05 ***	2.26
extrapos: V_comp	-7.423e-01	1.581e-01	-4.694	2.68e-06 ***	0.47
extrapos: N_comp	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&	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	

<b>Predictor</b>	<b><math>\beta</math></b>	<b>ASE</b>	<b>z-value</b>	<b>Pr(&gt; z )</b>	<b>O.R.</b>
intercept	-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:pass_zijn	2.057e+00	2.535e-01	8.113	4.94e-16 ***	7.82
Aux:worden	2.462e+00	2.296e-01	10.722	< 2e-16 ***	11.73
Aux:unclass	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:sep	1.352e+00	1.821e-01	7.426	1.12e-13 ***	3.87
mid_p: 3-5	7.083e-01	1.459e-01	4.856	1.20e-06 ***	2.03
mid_p: 6-8	8.288e-01	1.613e-01	5.139	2.76e-07 ***	2.29
mid_p: 9-11	8.270e-01	2.009e-01	4.116	3.85e-05 ***	2.29
mid_p: 12-14	9.446e-01	3.001e-01	3.147	0.001649 **	2.57
mid_p: >14	6.832e-01	3.653e-01	1.871	0.061407 .	1.98
synt_pers: none	5.404e-01	1.339e-01	4.035	5.47e-05 ***	1.72
pre-eindgroep: red	1.190e+00	1.363e-01	8.730	< 2e-16 ***	3.28
informat.: intermed.	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
bigram: sign	8.163e-01	1.850e-01	4.412	1.02e-05 ***	2.26
extrapos: V_comp	-7.423e-01	1.581e-01	-4.694	2.68e-06 ***	0.47
extrapos: N_comp	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&	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	

<b>Predictor</b>	<b><math>\beta</math></b>	<b>ASE</b>	<b>z-value</b>	<b>Pr(&gt; z )</b>	<b>O.R.</b>
intercept	-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:pass_zijn	2.057e+00	2.535e-01	8.113	4.94e-16 ***	7.82
Aux:worden	2.462e+00	2.296e-01	10.722	< 2e-16 ***	11.73
Aux:unclass	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:sep	1.352e+00	1.821e-01	7.426	1.12e-13 ***	3.87
mid_p: 3-5	7.083e-01	1.459e-01	4.856	1.20e-06 ***	2.03
mid_p: 6-8	8.288e-01	1.613e-01	5.139	2.76e-07 ***	2.29
mid_p: 9-11	8.270e-01	2.009e-01	4.116	3.85e-05 ***	2.29
mid_p: 12-14	9.446e-01	3.001e-01	3.147	0.001649 **	2.57
mid_p: >14	6.832e-01	3.653e-01	1.871	0.061407 .	1.98
synt_pers: none	5.404e-01	1.339e-01	4.035	5.47e-05 ***	1.72
pre-eindgroep: red	1.190e+00	1.363e-01	8.730	< 2e-16 ***	3.28
informat.: intermed.	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
bigram: sign	8.163e-01	1.850e-01	4.412	1.02e-05 ***	2.26
extrapos: V_comp	-7.423e-01	1.581e-01	-4.694	2.68e-06 ***	0.47

Predictor	$\beta$	ASE	z-value	Pr(> z )	O.R.
intercept	-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:pass_zijn	2.057e+00	2.535e-01	8.113	4.94e-16 ***	7.82
Aux:worden	2.462e+00	2.296e-01	10.722	< 2e-16 ***	11.73
Aux:unclass	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:sep	1.352e+00	1.821e-01	7.426	1.12e-13 ***	3.87
mid_p: 3-5	7.083e-01	1.459e-01	4.856	1.20e-06 ***	2.03
mid_p: 6-8	8.288e-01	1.613e-01	5.139	2.76e-07 ***	2.29
mid_p: 9-11	8.270e-01	2.009e-01	4.116	3.85e-05 ***	2.29
mid_p: 12-14	9.446e-01	3.001e-01	3.147	0.001649 **	2.57
mid_p: >14	6.832e-01	3.653e-01	1.871	0.061407 .	1.98
synt_pers: none	5.404e-01	1.339e-01	4.035	5.47e-05 ***	1.72
pre-eindgroep: red	1.190e+00	1.363e-01	8.730	< 2e-16 ***	3.28
informat.: intermed.	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
bigram: sign	8.163e-01	1.850e-01	4.412	1.02e-05 ***	2.26
extrapos: V_comp	-7.423e-01	1.581e-01	-4.694	2.68e-06 ***	0.47

<b>Predictor</b>	<b><math>\beta</math></b>	<b>ASE</b>	<b>z-value</b>	<b>Pr(&gt; z )</b>	<b>O.R.</b>
intercept	-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:pass_zijn	2.057e+00	2.535e-01	8.113	4.94e-16 ***	7.82
Aux:worden	2.462e+00	2.296e-01	10.722	< 2e-16 ***	11.73
Aux:unclass	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:sep	1.352e+00	1.821e-01	7.426	1.12e-13 ***	3.87
mid_p: 3-5	7.083e-01	1.459e-01	4.856	1.20e-06 ***	2.03
mid_p: 6-8	8.288e-01	1.613e-01	5.139	2.76e-07 ***	2.29
mid_p: 9-11	8.270e-01	2.009e-01	4.116	3.85e-05 ***	2.29
mid_p: 12-14	9.446e-01	3.001e-01	3.147	0.001649 **	2.57
mid_p: >14	6.832e-01	3.653e-01	1.871	0.061407 .	1.98
synt_pers: none	5.404e-01	1.339e-01	4.035	5.47e-05 ***	1.72
pre-eindgroep: red	1.190e+00	1.363e-01	8.730	< 2e-16 ***	3.28
informat.: intermed.	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
bigram: sign	8.163e-01	1.850e-01	4.412	1.02e-05 ***	2.26
extrapos: V_comp	-7.423e-01	1.581e-01	-4.694	2.68e-06 ***	0.47

# Multivariate statistical analysis

- No lack of fit:  $p > .05$
- No indication for multicollinearity (V.I.F.)
- Descriptive power (compared to dummy model):  
3031.8  $\rightarrow$  2325.7 (deviance reduction = 706.1)
- Predictive power:  $c = 0.803$  (after 100 bootstrap repetitions)
- After deletion of influential observations ( $n = 24$ ):  $c = 0.822$

# Summing up

- A systematic-empirical and quantitative corpus analysis provides a reliable insight into the complexity of language use, i.c. the choice between grammatical alternatives:
  - Several factors determine choice of word order *simultaneously*
  - Some factors have a higher impact than others
  - Most of the variation can be described adequately by the model
  - Most of the variation can be predicted by the model

# Multivariate nature of grammatical variation phenomena

- Within QLVL
  - Kris Heylen: order of verb arguments in German middle piece
  - Jose Tummers: inflection of prenominal adjectives in Dutch
  - Stefan Grondelaers: presentative *er* in Dutch
- Outside QLVL
  - Stefan Gries: particle placement
  - Stefanie Wulff: order of prenominal adjectives
  - Holger Diessel: order of main and adverbial clauses
  - ...

# Challenges for future research

- If grammar is multivariate in nature:
  - What are the implications for a model of grammar?
  - What are the general principles underlying grammatical variation?

# Challenges for future research

- If grammar is multivariate in nature:
  - What are the implications for a model of grammar?
  - What are the general principles underlying grammatical variation?

development of a multivariate model of grammar in which different types of functional features (structural/semantic, textual, lectal) co-determine the presence of linguistic features

# Challenges for future research

- If grammar is multivariate in nature:
  - What are the implications for a model of grammar?
  - What are the general principles underlying grammatical variation?

Statistical and linguistic variable reduction in order to identify higher-ordered principles

# Challenges for future research

- If grammar is multivariate in nature:
  - What are the implications for a model of grammar?
  - What are the general principles underlying grammatical variation?
- Towards a fully-fledged theoretical and methodological framework for usage-based grammatical variation research:
  - More multivariate analyses of grammatical variation (both intraconstituent and interconstituent variation)
  - A more principled selection of exp. variables
  - More attention to operationalisation of exp. variables



For further information:

[gert.desutter@arts.kuleuven.be](mailto:gert.desutter@arts.kuleuven.be)

<http://www.ling.arts.kuleuven.ac.be/qlvl/>