University of Amsterdam

Amsterdam Center for Language and Communication

Large-scale analysis of order variation in Dutch verbal clusters

Jelke Bloem, Arjen Versloot

Verbal clusters

- A word order variation in Dutch:
  1. ik denk dat ik het begrepen heb

     I think that I it understood have
  2. ik denk dat ik het heb begrepen

     I think that I it have understood
- Frisian, German: Only green order

Variation within grammar

- Some variation cannot be captured by rules
- Grammatical variation phenomena are multivariate in nature
  - If there was one cause we'd just have a rule
- On what basis do we choose between the two orders?

Explaining the variation (Coussé et al, 2008)

- (Regional) linguistic background
  - Single speaker variation?
- Mode of communication
- **Semantic factor**
- Discourse factor
  - Priming by previous syntactic structures

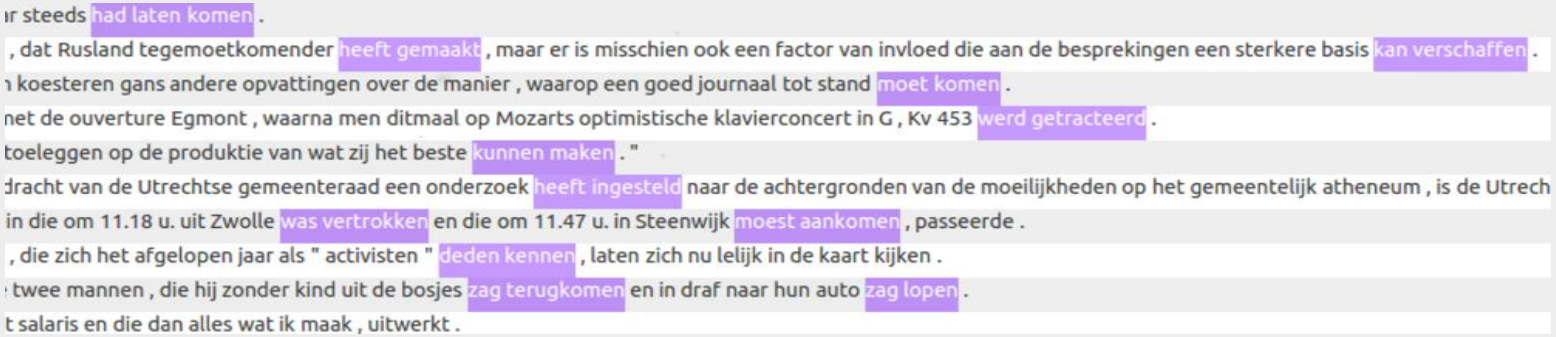
Corpus study (de Sutter, 2009)

- "De Standaard" part of CONDIV corpus
- Controlled for regional, register and diachronic variation
- Strict cluster criteria:
  - Only 'zijn' 'hebben' and 'worden' auxiliaries
  - Only complement clauses with 'dat' (that)
- Multivariate logistic regression model (10 variables)
- 2.390 manually verified clusters, 66.99% red order

Large-scale analysis

- Too limited definition of ‘verbal cluster’ by de Sutter (2009)
  - Unnecessary in a multivariate model
- Can be scaled up using large, automatically annotated corpora
  - Larger sample size
  - Coverage of more cluster types

Automatically annotated corpus



- Wikipedia part of “Lassy Large” corpus
- 137.093 clusters, 71.04% red order
- Syntactic annotation lets us formally define various types of clusters using DACT
- Limited to existing annotation
- May contain errors

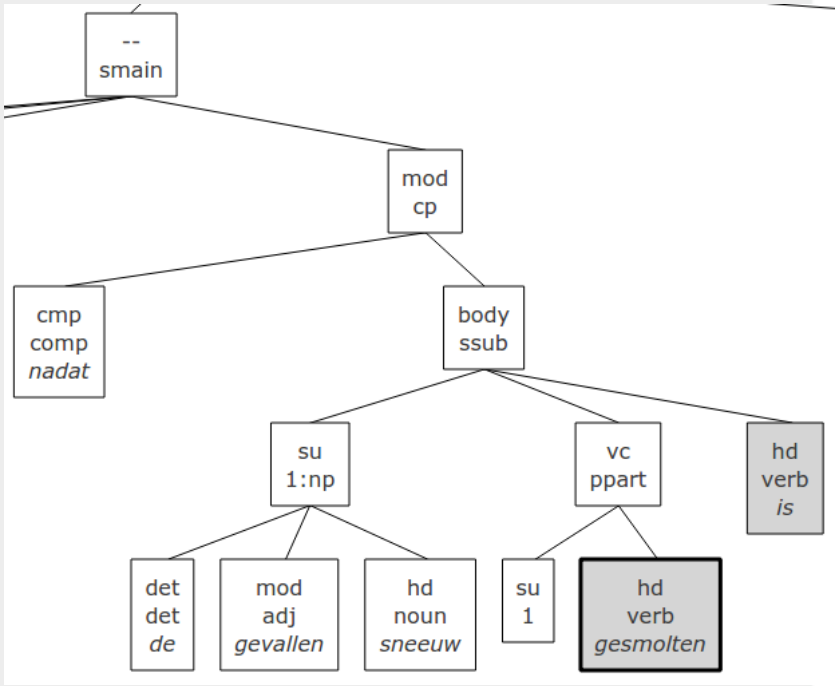
begrepen heb | heb begrepen

Variables (de Sutter, 2009)

- Accented syllable distance … naar hun **au**to is ge**lo**pen
- Separable main verb … heeft **af**gewassen (has washed **up**)
- Constituent after cluster … heeft gezien **dat het gebeurde**
- Length of the middle field … dat [hij naar hun auto] is gelopen
- Type of auxiliary copular-*zijn*/passive-*zijn*/time/*worden*
- Syntactic persistence …afgewassen heeft en …weggelopen is
- Main verb frequency … naar hun auto is **gelopen**
- Pre-verbal constituent: Informativity and inherence

Variables in Lassy Large corpus

- ~~Accented syllable distance~~
- Separable main verb
- Constituent after cluster
- Length of the middle field
- Type of auxiliary
- ~~Syntactic persistence~~
- Main verb frequency
- Pre-verbal constituent: Informativity and inherence



Additional cluster types

- Main clause
  - Rechters kunnen in principe niet worden ontslagen.

    Judges can in principle not be dismissed.
- Infinitival clusters
  - ... waardoor de reclame weer op tv **te** zien was

    … thus the ad again on tv to see was
- Aux/mod *worden hebben zijn / kunnen zullen willen laten mogen moeten blijven hoeven*
  - … dat iedereen hem ongestraft doden mocht

    … that everyone him with impunity kill **may**

Model comparison

Feature	De Sutter (2009)	This study
Separable main verb	3.87	4.34
Constituent after cluster: Complement of main verb Complement of preverbal noun Comp. or adjunct of main verb Comp. or adjunct of preverbal N.	Baseline: None 0.47 1.21 - -	Baseline: None - - 49.65 0.46
Length of middle field 3-5 words 6-8 words 9-11 words 12-14 words >14 words	Baseline: 0-2 words 2.03 2.29 2.29 2.57 1.98	Baseline: 0-2 words 2.22 2.90 3.00 3.04 2.85

The values are **odds ratios**, measuring effect size

OR=2.00 means: if this feature is present instead of the baseline, red order is 2 times more probable

Model comparison

Feature	De Sutter (2009)	This study
Type of auxiliary	Baseline: copular *zijn*	Baseline: *zijn*
Auxiliary of time	18.30	-
Passive *zijn*	7.82	-
worden	11.73	1.16
hebben	-	2.09
modal	-	140.72
Main verb frequency:	(from CELEX)	(from Lassy Large)
β	2.44^{E-06}	3.49^{E-08}
Std. Error	7.74^{E-07}**	1.74^{E-09}***
Inherence	2.26	1.93
Information value	Baseline: low	Baseline: low
Intermediate	1.41	1.36
High	1.94	1.22

Additional features

Feature	De Sutter (2009)	This study
Infinitival clusters	-	0.04
Main clause clusters	-	0.36

- Red infinitival (but only with *hoeven*)

... zodat de machinist niet in de locomotief zelf hoeft **te** zijn

… so that the operator not in the locomotive itself need to be

- Red main clause

Rechters kunnen in principe niet worden ontslagen.

Judges can in principle not be dismissed.

Similar model

Feature	Full model	Only aux/sub/fin
Type of auxiliary *worden* *hebben* modal	Baseline: *zijn* 1.16 2.09 140.72	Baseline: *zijn* 1.27 2.41 -
Constituent after cluster: Comp. or adjunct of main verb Comp. or adjunct of preverbal N.	Baseline: None 49.65 0.46	Baseline: None 59.33 0.48
Information value Intermediate High	Baseline: low 1.36 1.22	Baseline: low 1.10 1.03

Was it because our model includes more cluster types?

- No.

begrepen heb | heb begrepen

Replication

- Effect sizes largely similar
- Variables hold within a bigger model
- Cluster order is more or less affected by all 7 variables
- Some variables could not be measured

Main clause clusters

i.e. *Rechters kunnen in principe niet worden ontslagen.*

- 18.238 clusters, 57.42% red order
- Fewer modal clusters used?
  - 15.97% modal (subordinate clauses 25.54%)
  - Main clause OR: 0.35 over auxiliary clusters only
- Other differences with sub-clause clusters
  - Length of the middle field has smaller effect
  - Auxiliary type(*worden*, *hebben)* interact with main clause

Infinitival clusters

i.e. ... *waardoor de reclame weer op tv* ***te*** *zien was*

Feature	Inf cluster model	Full model
Type of auxiliary		Baseline: *zijn*
worden	0.40	1.16
hebben	0.57	2.09
modal	7762.22	140.72
Information value	Baseline: low	Baseline: low
Intermediate	3471.65	1.36
High	264.27	1.22
Inherence	64.91	1.93

- Novel findings (probably)

begrepen heb | heb begrepen

Dutch Europarl corpus

as part of Lassy Large corpus

- European Parliament proceedings texts
- 138.304 clusters, 86.78% red order!
- Variable effects largely similar

Feature	Europarl model	Wiki model
worden *hebben* modal	1.62 2.57 323.46	1.16 2.09 140.72
Comp. or adjunct of main verb Comp. or adjunct of preverbal noun	31.22 0.46	49.65 0.46

Semantic factor: Collostructional analysis

(Stefanowitsch & Gries, 2003)

- Relationship between a construction (red/green) and the words that fill its slots

… dat ik het **begrepen** heb
… dat ik het **gezien** heb
… dat ik het **gehoord** heb
… dat ik het **geschopt** heb
…

… dat ik het heb **gemaakt**
… dat ik het heb **bedacht**
… dat ik het heb **gehoord**
… dat ik het heb **beschreven**
…

- Calculate most strongly associated **collexemes**
  - Fisher's Exact Test

Collostructional analysis

Auxiliary, subordinate clause clusters only, cutoff=15

Main verbs - Odds ratio - Red - Green

```
1 --- afkondigen   inf    29    -
2 --- neerzetten   inf    24    -
3 --- uitmaken     inf    21    -
4 --- aanhouden    inf    21    -
5 --- optekenen    inf    19    -
6 --- overgeven    inf    18    -
7 --- aanschaffen  inf    17    -
8 --- uitschrijven inf    16    -
9 --- plaatsvinden 33.34  182   3
10--- indienen     22.95  42    1
```

Pattern?

Collostructional analysis

Auxiliary, subordinate clause clusters only, cutoff=100, no particle verbs

Main verbs	Odds ratio	Red	Green
1 --- staan	7.81	583	51
2 --- gaan	6.74	751	76
3 --- hebben	6.40	882	94
4 --- zitten	5.70	200	24
5 --- zijn	5.50	2583	317
6 --- waarnemen	5.32	233	30
7 --- ondergaan	5.11	179	24
8 --- gooien	5.06	133	18
9 --- blijven	4.68	748	109
10--- geworden	4.42	2509	383

Main verbs	Odds ratio	Red	Green
1 --- verplichten	20.44	13	182
2 --- zien	17.36	148	1751
3 --- danken	14.02	20	288
4 --- vinden	13.96	87	830
5 --- herkennen	7.08	20	97
6 --- relateren	6.70	22	101
7 --- huwen	5.94	28	114
8 --- besmetten	4.87	24	80
9 --- wijten	4.33	32	95
10--- bestemmen	4.28	61	179

Stative verbs? Perception verbs? Animate verbs?

Frequency effect?

begrepen heb | heb begrepen

Conclusions

- Replicated and extended a linguistic study using automatically annotated corpus
- Comprehensive model of Dutch verbal clusters
- Automatic approach is easily extended
  - Study regional/register/diachronic variation
- de Sutter (2009)'s variables generalize to another domain
- Larger sample allows more detailed analysis

Discussion

- On what basis do we choose between the two orders?
- Examine unexpected effects
- Explore semantics intuition with collostructional analysis
- Test whether factors can be dropped or reformulated
- Add interactions between variables

References

Coussé, E., Arfs, M., & De Sutter, G. (2008). Variabele werkwoordsvolgorde in de Nederlandse werkwoordelijke eindgroep. Een taalgebruiksgebaseerd perspectief op de synchronie en diachronie van de zgn. rode en groene woordvolgorde. In G. Rawoens (Ed.), *Taal aan den lijve. Het gebruik van corpora in taalkundig onderzoek en taalonderwijs* (pp. 29–47). Gent: Academia Press.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics*, *8*(2), 209-243.

Sutter, G. D. (2009). Towards a multivariate model of grammar: The case of word order variation in Dutch clause final verb clusters. *Describing and modeling variation in grammar*, *204*, 225-254.

- Thanks to Fred Weerman for comments