University of Amsterdam

Amsterdam Center for Language and Communication

Applying automatically parsed corpora to the study of language variation

Jelke Bloem, Arjen Versloot, Fred Weerman

Language variation

- Grammars often contain optionality
  - Same meaning, different form
- Grammatical variation phenomena are multivariate in nature
  - Rules can only outline the variation
- On what basis do we choose between the options?

Dative alternation

- A word order variation in English:
  1. He gave [**his friend**] [**the ticket**]
  2. He gave [**the ticket**] to [**his friend**]
- No simple rule on when to use one or the other
- Probabilistically modeled using 14 variables
  i.e. animacy of recipient, pronominality of recipient, given-ness (Bresnan et al., 2007)
- Switchboard corpus (3M words, 2360 instances)

Automatically parsed corpora

- Fewer annotation resources required
  - More data
- Exact definition of construction
- Flexible
- Contains errors (‘random’ or systematic)
- Annotation may be a limiting factor

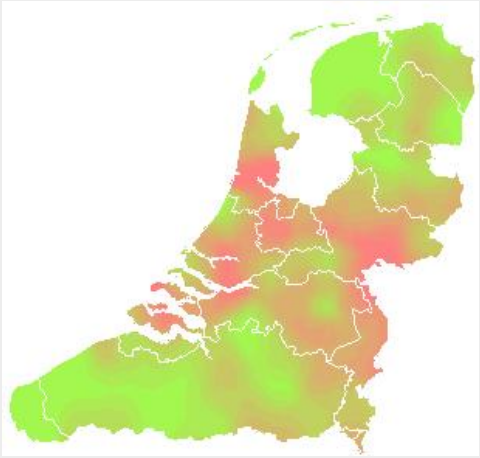
```
(@root="word" or @root="heb" or
@root="ben") and
(parent::node[( @cat="rel" or
@cat="ssub" or @cat="oti" or
@cat="cp" or @cat="svan" or
@cat="ahi")] or
parent::node[@rel="vc" and
parent::node[(@cat="smain" or
@cat="sv1")]] or
parent::node[@cat="inf" and
```

Case study: 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

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



understood have | have understood

Corpus study (de Sutter, 2009)

- "De Standaard" part of CONDIV corpus (3.2M words)
- Controlled for regional, register and diachronic variation
- Strict cluster criteria:
  - Only 'hebben' (*to have)* 'zijn' and 'worden' *(to be)* auxiliaries
  - Only complement clauses with 'dat' (*that*)
- Multivariate logistic regression model (10 variables)
- 2.390 manually verified clusters, 66.99% red order

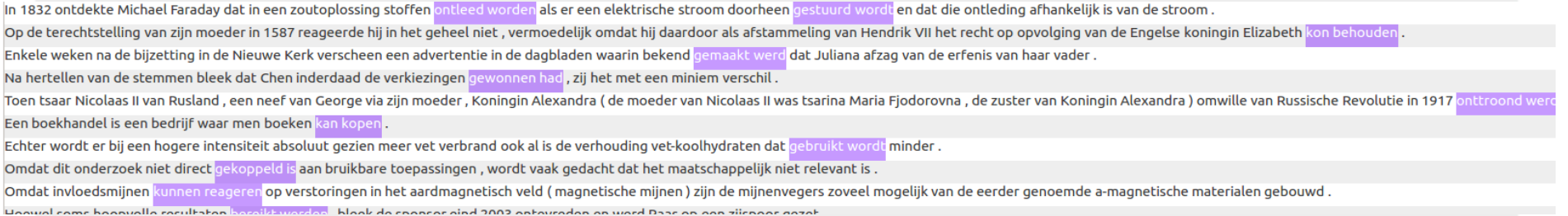
understood have | have understood

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

understood have | have understood

Automatically annotated corpus



- Wikipedia part of “Lassy Large” corpus
- 145M tokens, 411.623 clusters, 71.65% red order
- Syntactic annotation lets us formally define various types of clusters using DACT (X-path queries)
  - Dependency trees (+ features)
  - May contain errors: 88.38% parser accuracy

understood have | have understood

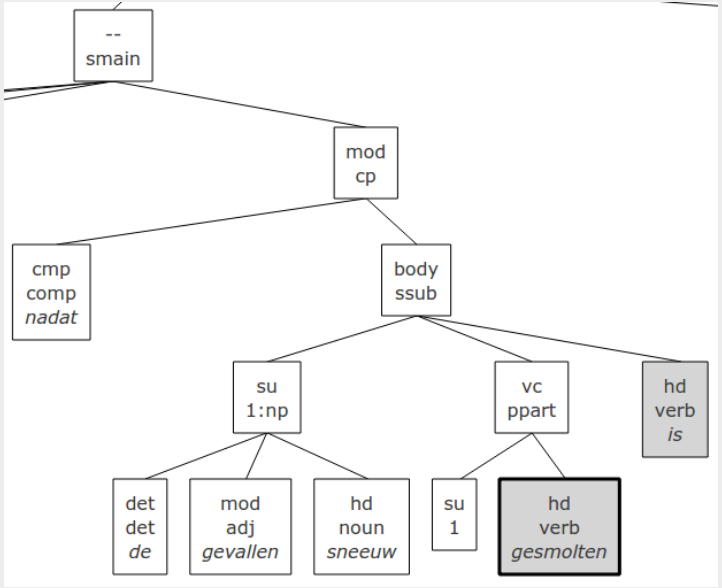
Variables of the original study (de Sutter, 2009)

- Accented syllable distance … naar hun **au**-to is ge-**lo**-pen **n=4**
- 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 **n=4**
- 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

understood have | have understood

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



understood have | have understood

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**

understood have | have understood

Model comparison

Feature	De Sutter (2009)	This study
Separable main verb		
No	1.00	1.00
Yes	3.87	4.92
Constituent after cluster		
None	1.00	1.00
Complement of main verb	0.47	-
Complement of preverbal noun	1.21	-
Comp. or adjunct of main verb	-	51.44
Comp. or adjunct of preverbal N.	-	0.44
Length of middle field		
0-2 words	1.00	1.00
3-5 words	2.03	2.42
6-8 words	2.29	3.23
9-11 words	2.29	3.34
12-14 words	2.57	3.33
>14 words	1.98	3.15

Feature	De Sutter (2009)	This study
Type of auxiliary		
Copular *zijn*	1.00	-
Auxiliary of time	18.30	-
Passive *zijn*	7.82	-
zijn	-	1.00
worden	11.73	1.19
hebben	-	2.19
modal	-	132.42
Main verb frequency	(from CELEX)	(Lassy Large)
β	2.44^{E-06}	3.73^{E-08}
Std. Error	7.74^{E-07}**	1.05^{E-08}***
Inherence		
No	1.00	1.00
Yes	2.26	2.10
Information value		
Low	1.00	1.00
Intermediate	1.41	1.21
High	1.94	1.11

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

understood have | have understood

Model predictive power

Concordance index c

Model	C-index	Nr. of features	Data
De Sutter (2009)	0.8030	10	AUX/Sub only
Full model	0.8635	9	All clusters
Small model	0.7649	7	AUX/Sub only

Full model intercept = 0.6035

* Values actually not directly comparable

* c=0.5 is chance level, but the gold standard is not 1…

understood have | have understood

Stepwise regression

- Minimize Akaike Information Criterion (AIC)
- Indicates relative importance of the features

Feature	AIC
0. <none>	490828
1. Type of auxiliary	413913
2. Constituent after cluster	349852
3. Finiteness	338758
4. Length middle field	332781
5. Clause type	325857
6. Frequency main verb	324371
7. Inherence	323201
8. Separable verb	322519
9. Information value	322000

understood have | have understood

Replication summary

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

understood have | have understood

Additional features

Feature	De Sutter (2009)	This study
Infinitival clusters		
No	-	1.00
Yes	-	0.03
Clause type		
Subordinate clause	-	1.00
Main clause	-	0.34

- Red pure-infinitival cluster (but only with *hoeven* 'need')

  ... 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 principe not be dismissed.

understood have | have understood

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
zijn	1.00	1.00
worden (to be)	1.62	1.19
hebben (to have)	2.57	2.19
modal	323.46	132.42
None	1.00	1.00
Comp. or adjunct of main verb	31.22	51.44
Comp. or adjunct of preverbal noun	0.46	0.44

understood have | have understood

Semantic factor: Collostructional analysis

(Stefanowitsch & Gries, 2003)

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

*… that I it **VERB** have*
… dat ik het **begrepen** heb
… dat ik het **gezien** heb
… dat ik het **gehoord** heb
… dat ik het **geschopt** heb
…

*… that I it have **VERB***
… 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 or other association measure

understood have | have understood

Collostructional analysis

Auxiliary, subordinate clause clusters only, cutoff=15

Main verbs		Odds ratio	Red	Green
1 --- afkondigen	(proclaim)	inf	29	0
2 --- neerzetten	(put down)	inf	24	0
3 --- uitmaken	(make out)	inf	21	0
4 --- aanhouden	(keep on)	inf	21	0
5 --- optekenen	(draw up)	inf	19	0
6 --- overgeven	(surrender / throw up)	inf	18	0
7 --- aanschaffen	(purchase)	inf	17	0
8 --- uitschrijven	(unsubscribe)	inf	16	0
9 --- plaatsvinden	(take place)	33.34	182	3
10--- indienen	(send in)	22.95	42	1

Pattern?

understood have | have understood

Collostructional analysis

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

Main verbs		Odds ratio	Red	Green
1 -- verplichten (to oblige)	ST	20.44	13	182
2 -- zien (to see)	ST	17.36	148	1751
3 -- danken (to thank)		14.02	20	288
4 -- vinden (to find)	ST	13.96	87	830
5 -- herkennen (to recognize)	ST	7.08	20	97
6 -- relateren (to relate)	ST	6.70	22	101
7 -- huwen (to marry)		5.94	28	11
8 -- besmetten (to infect)		4.87	24	80
9 -- wijten (to blame)		4.33	32	95
10-- bestemmen (to assign)		4.28	61	179

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

Patterns: Semantic classes? Stative/dynamic verbs?

understood have | have understood

Conclusions

- Replicated and extended a linguistic study using an 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

- Extend further:
  - Corpus of Spoken Dutch
  - A corpus with author/region/time metadata
  - Larger verbal clusters
- Semantics: more can be done
- Automatically annotated corpora for:
  - Dative alternation
  - ‘that’-optionality
  - Any other probabilistic phenomenon
- New types of corpora as NLP tools get better

understood have | have understood

References

J. Bresnan, A. Cueni, T. Nikitina, R. H. Baayen, et al. Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94, 2007.

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.