University of Amsterdam

Amsterdam Center for Language and Communication

Using automatically annotated corpora in language variation research

Jelke Bloem, Arjen Versloot, Fred Weerman

Language variation

- Grammars often contain optionality
  - Same meaning, different form
    1. He said [**that**] he would do it
- 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)

```
(@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
```

Automatically parsed corpora

- Fewer annotation resources required
  - Dutch LASSY Small corpus: 1M tokens
  - Dutch LASSY Large corpus: 700M tokens
- Flexible
- Exact definition of construction
- Contains errors (‘random’ or systematic)
- Annotation may constrain what can be researched

Case study: Dutch 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

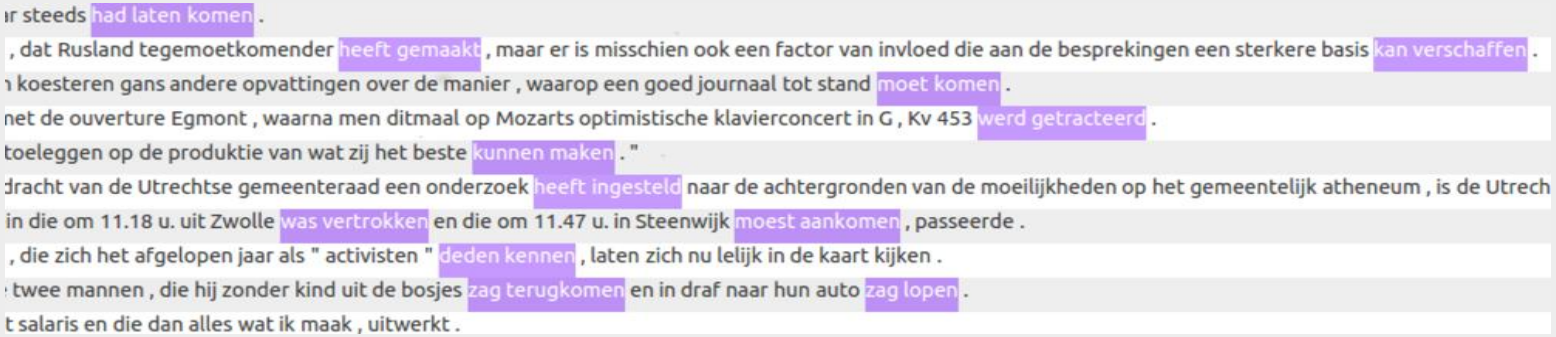
Manual corpus study (de Sutter, 2009)

- "De Standaard" part of CONDIV corpus (3.2M words)
- Controlled for regional, register and diachronic variation, specific cluster types

Select data -> Semi-automatically retrieve data -> Manually verify results

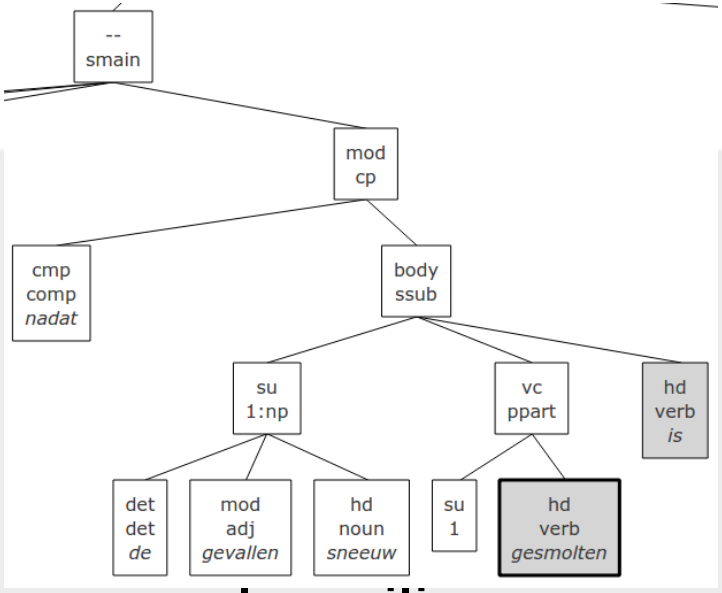
- Multivariate logistic regression model (10 variables)
- 2.390 manually verified clusters, 66.99% red order

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)
- Limited to existing annotation
- May contain errors: 88.38% parser accuracy

understood have | have understood



Automatic annotation

Limitations of the annotation:

- Accented syllable distance
- 'to be' = passive, copulative, temporal auxiliary…

Workflow: Choose corpus -> Define searches-> Automatically retrieve sentences -> Automatically extract features

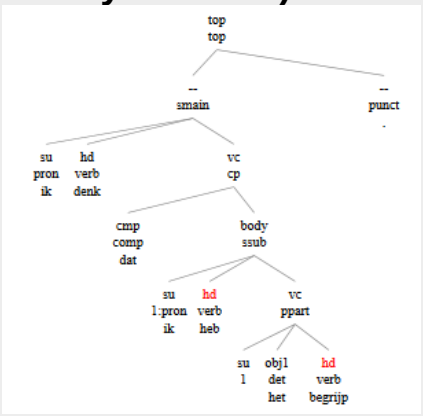
```
//node[@lemma="hebben" and (some $x in //node[@pt="ww"]
   satisfies (number(@begin) = $x/number(@end)))]
```

understood have | have understood

Using GrETEL (Augustinus, Vandeghinste, and Van Eynde 2012)

- Example-based treebank querying:

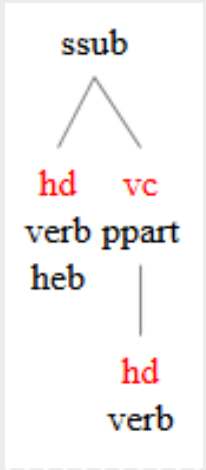
ik denk dat ik het heb begrepen



sentence	Ik	denk	dat	ik	het	heb	begrepen	.
word	○	○	○	○	○	○	○	○
lemma	○	○	○	○	○	◉	○	○
word class	○	○	○	○	○	○	◉	○
optional in search	◉	◉	◉	◉	◉	○	○	◉

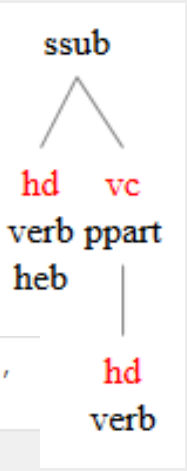
OPTIONS

☑ Respect word order



understood have | have understood

Using GrETEL: Output



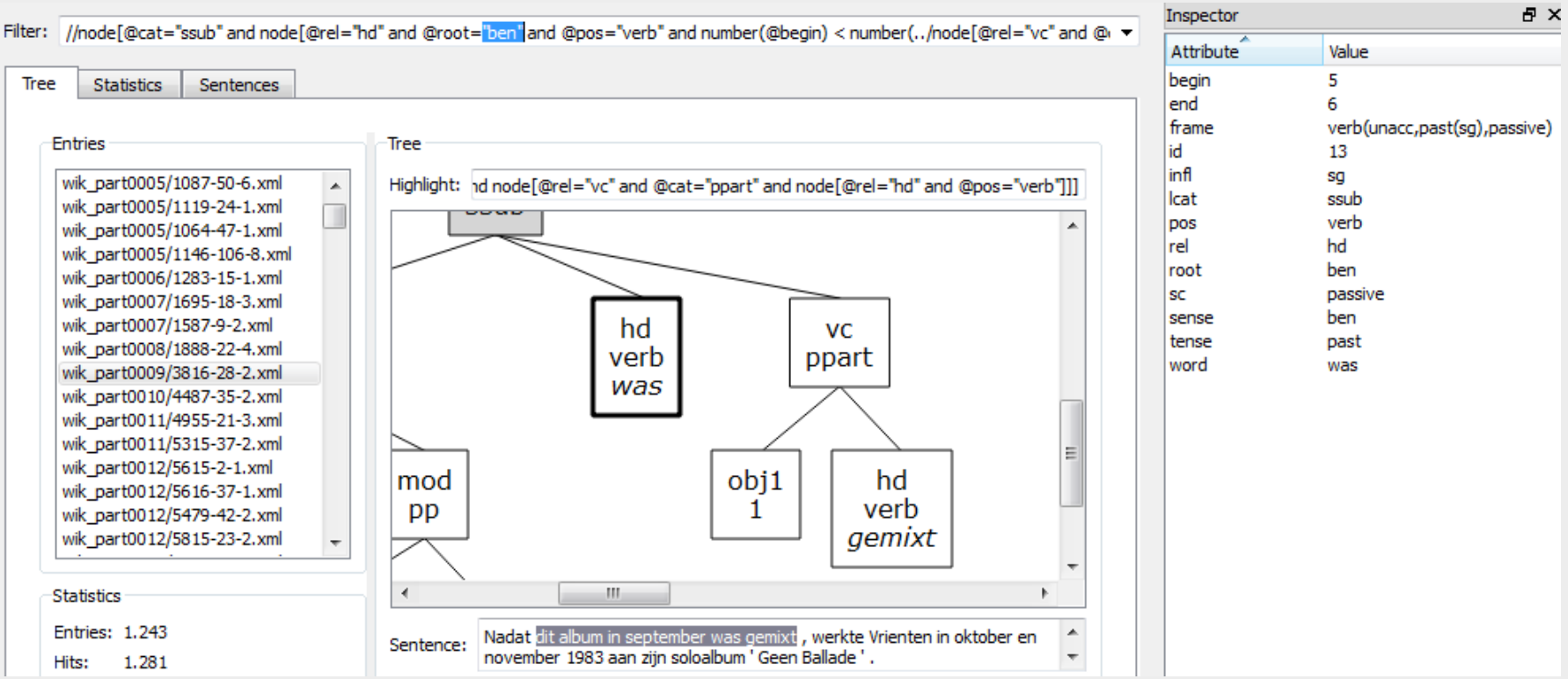
WR-P-P-I-0000000060.p.2.s.3	Als patiënten aandringen op antibiotica , of als **artsen eerder een longontsteking hadden gemist** , werd er gemakkelijker voorgeschreven .
WR-P-P-I-0000000249.p.13.s.3	Nu bovendien blijkt dat **bedrijfsrevisor Andersen Consulting wellicht verschillende bedrijfsdocumenten en e-mails heeft vernietigd** , wordt een gerechtelijk onderzoek ingesteld .
WS-U-E-A-0000000017.p.37.s.3	Toch waren er minder deelnemers dan verwacht , waarschijnlijk doordat **belangrijke moslimorganisaties de mensen hadden afgeraden om mee te doen** .
WR-P-E-I-0000049645.p.1.s.128.2	In 1926 , toen de frank ten opzichte van bijvoorbeeld de dollar (10 $ toen was 101 $ waard in 2002) op vier jaar tijd tot een derde in waarde gedaald was en nadat **België eindelijk een buitenlandse lening had losgepeuterd** , voerde minister van financiën Janssen een hervorming door om de frank te stabiliseren .
WR-P-P-H-0000000058.p.5.s.1	Feit is dat **Bellaart van het vernieuwde en verjongde team nog geen swingende eenheid heeft gemaakt** .

```
//node[@cat="ssub" and node[@rel="hd" and @root="heb" and @pos="verb" and
    number(@begin) < number(../node[@rel="vc" and @cat="ppart"]/node[@rel="hd"
    and @pos="verb"]/@begin)] and node[@rel="vc" and @cat="ppart" and
    node[@rel="hd" and @pos="verb"]]]
```

understood have | have understood

Using DACT: Query-based… querying

De Kok (2010)



Same query syntax: Insert the query that GrETEL produces

understood have | have understood

Verbal cluster study: Results

- 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

Results: 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
* The gold standard is not 1…

understood have | have understood

Large-scale: Collostructional analysis

(Stefanowitsch & Gries, 2003)

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

*… that I it **VERB** have*

Main verbs		Odds ratio	Red	Green
1	verplichten (to oblige)	20.44	13	182
2	zien (to see)	17.36	148	1751
3	danken (to thank)	14.02	20	288
4	vinden (to find)	13.96	87	830
5	herkennen (to recognize)	7.08	20	97

*… that I it have **VERB***

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

understood have | have understood

Verbal cluster study summary

- Variable effects largely similar to previous work
- Variables hold within a bigger model
- Variables hold in other domain: Europarl corpus
- All variables in the study are associated with cluster word order
- Some variables could not be measured
- Detailed results on our poster
  or in Bloem, Versloot & Weerman (2014)

Conclusions

- Automatically annotated corpora are particularly useful for language variation studies
- Replicated and extended a manual linguistic study
- Larger sample allows more detailed analysis
- Automatic approach is easily extended
  - Study regional/register/diachronic variation
- Example-based querying, standard query syntax and standard annotation format help accessibility

Discussion

- Automatically annotated corpora for
  - Dative alternation
  - 'that'-optionality
  - Any other probabilistic phenomenon
- Extend the study:
  - Corpus of Spoken Dutch
  - A corpus with writer/region/time metadata
  - Larger verbal clusters
- New types of corpora as NLP tools get better

References

L, Augustinus, V. Vandeghinste, and F. Van Eynde (2012). Example-Based Treebank Querying. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012).* Istanbul, Turkey.

J. Bloem, A.P. Versloot & F.P. Weerman (2014). Applying automatically parsed corpora to the study of language variation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1974-1984). Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

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

GrETEL: http://nederbooms.ccl.kuleuven.be/eng/gretel

DACT: http://rug-compling.github.io/dact/