Cues for verb-finality in adult-directed and child-directed Dutch

Aleksei Nazarov

Utrecht University

Institute for Language Studies

Overview

- Dutch PP and separable verb particle (SVP) order w.r.t. verb
  - Optionality & difference in transparency to learning Dutch basic word order
- GrETEL: existing tool to facilitate searching treebanks
- Child-directed and adult-directed texts searched
  - Spoken and written
- Results:
  - PP order tends to be transparent, and more so in child-directed speech
  - SVP order tends to be opaque, but less so in child-directed speech
  - No major threats to learnability expected from existing results

Dutch word order

- Dutch generally acknowledged as verb-final language (e.g. Koster 1975)
  - Verb-finality in embedded clauses, but V2 in main clauses

Die doceerde taalkunde
they taught linguistics
*… dat die doceerde taalkunde
that they taught linguistics

*Die taalkunde doceerde
they linguistics taught
… dat die taalkunde doceerde
that they linguistics taught

- Creates learnability challenge: main clauses are generally more frequent, but only embedded clauses show “underlying” word order
  - Constructions with non-verb-final embedded clauses: additional problem?

Dutch PP constructions (Broekhuis & Corver 2020a,b)

- In main clauses, PPs can be only after the verb
- Embedded clauses: PPs before or after the verb

Die doceerde in Utrecht
they taught in Utrecht

… dat die doceerde in Utrecht
that they taught in Utrecht

*Die in Utrecht doceerde
they in Utrecht taught

… dat die in Utrecht doceerde
that they in Utrecht taught

- Postverbal PP in embedded clauses: violates strict verb-finality, potentially takes away from cues (*non-transparent* word order)

Dutch particles

- Separable verb particles (SVPs; Booij & Audring 2020):
  - Separated from verb in main clauses

Die gaat in Utrecht door-t_{gaat} *Die door-gaat in Utrecht $t_{door\text{-}gaat}$
they goes in Utrecht SVP   *they SVP-goes in Utrecht*
"They (will) continue in Utrecht"

  - In verb clusters, may be adjacent to or separate from their host verb
  - Separate word order provides a cue for verb-finality (*transparent*)

adjacent: … dat die [[[$t_{door\text{-}gaan}$] t_{willen}] heeft] willen door-gaan
that they has want.INF SVP-continue.INF
separate: … dat die [[[door-t_{gaan}] t_{willen}] heeft] willen gaan
that they SVP has want.INF continue.INF

Combination

- SVPs and PPs combined:
  - Hard constraint: when SVP and PP are adjacent, PP comes first

    … dat die in Utrecht door heeft willen gaan
    that they in Utrecht SVP has want continue
    *… dat die door in Utrecht heeft willen gaan
    that they SVP in Utrecht has want continue

  - Otherwise, SVP and PP order seems independent

    in U door heeft willen gaan / door heeft willen gaan in U
    in U heeft willen doorgaan / **heeft willen doorgaan in U**

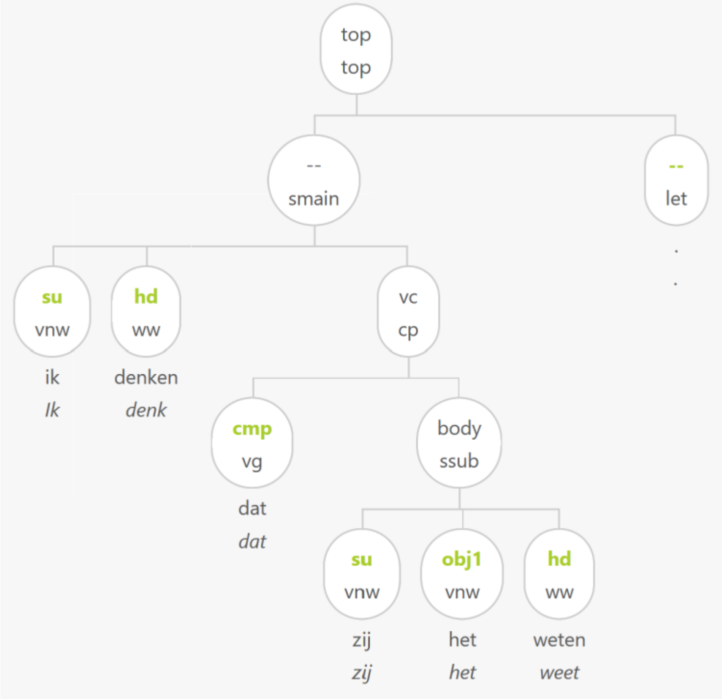
    - How often do *doubly opaque* orders occur? Tradeoff between SVP/PP?

GrETEL

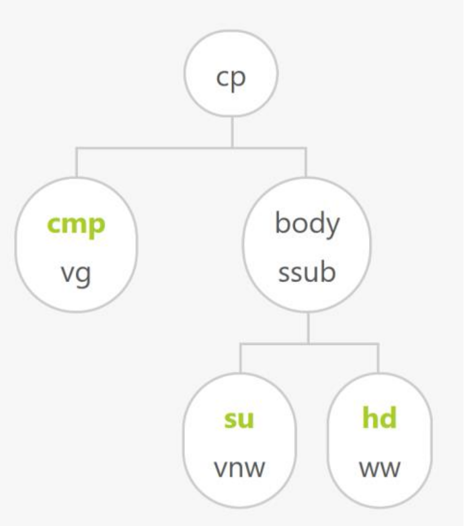
- Greedy Extraction of Trees for Empirical Linguistics, version 4 used (Odijk et al. 2018)
- Parses corpora with Alpino parser (Bouma et al. 2001)
  - Trees with constituency and dependency elements
  - Pre-uploaded corpora, some of which have been checked manually
- Searches Alpino-parsed treebanks (XML) using XPath
- Crucially: generates XPath expressions by generalizing from example sentences
  - Give the tool an example sentence and ask it to find sentences/constructions “just like this”

GrETEL

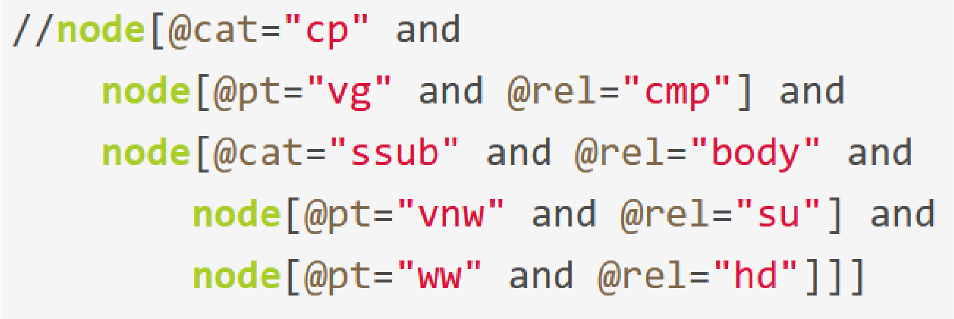
- Example: “Ik denk dat zij het weet.”
  I think that she it knows
- Parse by Alpino =>



- Select only word class + features of “dat”, “zij”, “weet”:



- Resulting query:



```
//node[@cat="cp" and
    node[@pt="vg" and @rel="cmp"] and
    node[@cat="ssub" and @rel="body" and
        node[@pt="vnw" and @rel="su"] and
        node[@pt="ww" and @rel="hd"]]]
```

Corpora searched

- LASSY Klein (Noord et al. 2013): **written, adult-directed**;
  parses manually checked
- CGN (Corpus Gesproken Nederlands, Van Eerten 2007):
  spoken, adult-directed;
  parses manually checked
- BasiLex (Tellings et al. 2015): **written, child-directed** (textbooks for younger children; subset of corpus uploaded to GrETEL 4)
- Dutch subcorpora of CHILDES (MacWhinney 2000): **spoken, child-directed** + child productions

- Queries developed and calibrated initially on LASSY Klein corpus, then re-calibrated on CGN and other corpora

Statistics of construction occurrence

	Written				Spoken					
	Adult-directed (LASSY)		Child-directed (BasiLex)		Adult-directed (CGN)		Child-directed (CHILDES)		Child-produced (CHILDES)	
Phenomenon	abs	Per 1000s	abs	Per 1000s	abs	Per 1000s	abs	Per 1000s	abs	Per 1000s
PP+V	16,928	260	3,362	128	12,758	98	4,075	10	1,257	4
SVP+V	695	11	115	4	755	6	78	<1	27	<1
PP+SVP+V	280	4	37	1	218	2	11	<1	1	<1
Total number of sents	65,200		26,239		129,923		351,859		351,859	

- PPs in embedded clauses much more common than SVPs
- Both constructions occur more rarely towards right of the table (= in spoken or child-directed material)

PP order results

	Written		Spoken		
	Adult-directed	Child-directed	Adult-directed	Child-directed	Child-produced
% preverbal PP	51%	74%	71%	77%	80%
total PP sentences	16,928	3,362	12,758	4,075	1,368

- Preverbal PPs are expected to be the unmarked order (transparent wrt verb-finality)
- Proportionally more preverbal PPs towards right of table (spoken or child-directed)
- Child-directed and child-produced speech have similar rates (still sign. different)
    - Child-directed: slight positive correlation between child’s age and preverbal PPs

PP order: per PP function

	Written		**Spoken**		
	Adult-directed	Child-directed	Adult-directed	Child-directed	Child-produced
% preverbal PP, location/direction	79%	92%	88%	88%	86%
% preverbal PP, modification	**50%**	65%	64%	63%	70%
% preverbal PP, predicate complement	37%	70%	60%	**75%**	**77%**
% preverbal PP, overall	51%	74%	71%	77%	80%

- These 3 Alpino-coded PP functions account for 92%-98% of cases (as per corpus)
- Location/direction: overattested in preverbal position (O/E, Kendall’s tau)
- Modification, pred. comp.: underattested with preverbal PP, except bold numbers

SVP order results

	Written		Spoken		
	Adult-directed	Child-directed	Adult-directed	Child-directed	Child-produced
% separated SVP	12%	26%	30%	51%	48%
total SVP sentences	695	115	755	78	27

- Tendency against separated particles (even though they are more transparent)
  - But more separated particles towards right hand side of table
- Are separated particles dispreferred because of processing difficulties (non-adjacency)?
  - No significant correlation between length of verb cluster and separate/adjacent SVP (except in spoken adult-directed corpus)

Combination results (expected percentages)

	Written		Spoken		
	Adult-directed	Child-directed	Adult-directed	Child-directed	Child-produced
% PP Pcl… V	5% (6%)	32% (20%)	15% (21%)	46% (39%)	
% Pcl… V PP	6% (6%)	5% (7%)	9% (9%)	9% (12%)	100%
% PP Pcl-V	50% (45%)	46% (55%)	47% (50%)	27% (37%)	
% Pcl-V PP	40% (44%)	16% (19%)	29% (21%)	18% (11%)	
Total PP+Pcl+V	280	37	218	11	1

- No apparent interaction: no consistent over- or underattestation
  - Doubly opaque order (nr. 4) is not extremely rare
- Word order constraint obeyed in all token sentences

Discussion

- PPs: preverbal position preferred across the board, but even more so in child-directed speech
  - This enhances the evidence for word-finality, especially in child-directed speech
  - Child-directed and child-produced speech matches quite well
- SVPs: separated particles dispreferred across the board, but less so in child-directed speech
  - Number of SVP tokens is quite small, should be no significant obstacle for learnability
- The constructions don’t seem to interact
  - No tradeoff in terms of evidence for word-finality, *double opacity* allowed

Conclusion

- PP and SVP word order varies, but should not be a major problem for learning of verb-finality
  - PP word order is mostly transparent in child-directed speech (and child productions)
  - SVP word order is not always transparent, but few SVP tokens anyway
- GrETEL is effective in searching for infrequent constructions
- Future work:
  - Modelling learnability of verb finality
  - Which factors lead to less transparent orders, why more in adult-directed text?
  - Work with larger corpora

Acknowledgments

- Many thanks to Roberta d’Alessandro, Jelke Bloem, David Erschler, and especially Jan Odijk for their help with and advice on various part of this project, and to audiences at CLiN and WECOL.
- This project is supported by CLARIAH Grant
  CP-WP3-22-005
  The acquisition of word order in Dutch non-V2 clauses: PPs and verb particles

References

- Booij, Geert & Audring, Jenny. 2020. Separable complex verbs (SCVs). Taalportaal. Retrieved from https://taalportaal.org/taalportaal/topic/pid/topic-13998813296768009.
- Bouma, Gosse, Gertjan Van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37:45–59, 2001.
- Broekhuis, Hans, and Norbert Corver. 2020a. 10.1. Placement of the finite verb. *Taalportaal*.
- Broekhuis, Hans & Norbert Corver. 2020b. 12.3. Modifiers of the clause. *Taalportaal*.
- Eerten, Laura van. 2007. Over het Corpus Gesproken Nederlands. Nederlandse Taalkunde 12(3):194–215.
- Koster, Jan. 1975. Dutch as an SOV language. *Linguistic Analysis* 1:111-136.
- MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.
- Noord, Gertjan van, Gosse Bouma, Frank van Eynde, Daniël de Kok, Jelmer van den Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large Scale Syntactic Annotation of Written Dutch: Lassy. In. In Peter Spyns, and Jan Odijk (eds.), Essential Speech and Language Technology for Dutch: Results by the STEVIN programme, 147-164. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer.
- Odijk, Jan, Martijn van der Klis and Sheean Spoel. 2018. “Extensions to the GrETEL treebank query application” In: Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories. Prague, Czech Republic. pp. 46-55.
- Tellings, Agnes, Micha Hulsbosch, Anne Vermeer, and Antal van den Bosch. 2015. BasiLex: an 11.5-million words corpus of Dutch texts written for children. *Computational Linguistics in the Netherlands Journal* 4:191-208.