How can we determine at what level of abstraction lectal distinctions operate?

A case study of the alternation(s) between the Dutch direct and prepositional object

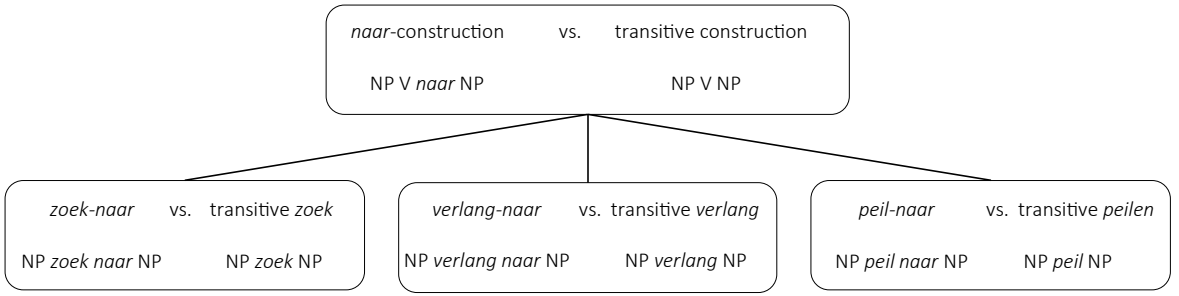
Dirk Pijpops, Dirk Speelman, Stefan Grondelaers & Freek Van de Velde

Research Foundation Flanders (FWO); QLVL, University of Leuven; V&D Radboud University of Nijmegen

(1) *Samen zoeken zij (naar) een oplossing*.
'Together, they are searching a solution'

(2) *Maar de politiek verlangt nu (naar) scherpere maatregelen* (...)
'But politicians now desire more severe measures (...)'

(3) *Margreet peilde (naar) hun reacties*.
'Margreet gauged their reactions.'



LECTAL HYPOTHESES

Belgian Dutch is more heterogeneous than Netherlandic Dutch

→ Models based on Belgian Dutch will have lower predictive quality

Variation is more strictly fixed by lexical biases and semantic distinctions in the Netherlands

→ Lexical and semantic factors should lead to a greater increase in predictive quality for Netherlandic models than for Belgian models.

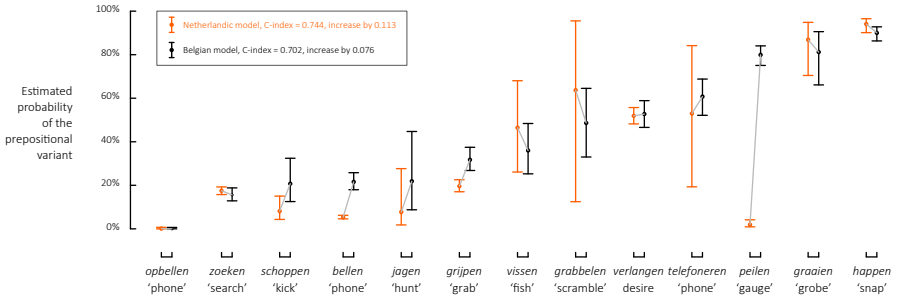
SEMANTIC & LEXICAL FACTORS

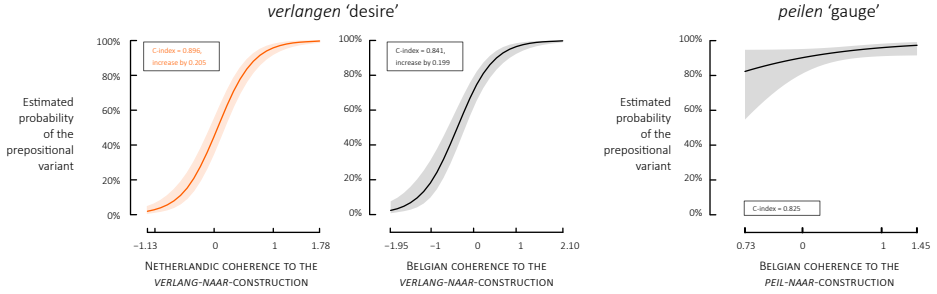
Lexical Origin Hypothesis: non-lexical constructions acquire their meaning from their prototypical slot fillers

→ Identify 5 most prototypical slot fillers: collostructional analyses

Principle of Semantic Coherence: lexemes will more often combine with argument structure constructions that are semantically compatible

→ Calculate semantic distance to the prototypical slot fillers: distributional vectors





Dirk Pijpops is currently finishing a PhD on variation in Dutch argument structure at the University of Leuven, supervised by Dirk Speelman, Stefan Grondelaers and Freek Van de Velde.

dirk.pijpops@kuleuven.be

Selecting the verbs

We investigate the alternation between the Dutch transitive construction and the prepositional intransitive construction. All analyses were executed on the Sonar corpus, annotated by the Alpino-parser (van Noord 2006; Oostdijk et al. 2013). The New-Media components of the corpus were not used, nor were the discussion lists. We charted out all verbs that can alternate as follows. First, all unique combinations of a verb and a preposition were selected that appeared both in the transitive and prepositional intransitive construction, and for which at least 3 different object roots, i.e. the root of the syntactic head of the theme argument, appeared in both constructions. Next, 4 annotators independently judged 650 of these combinations on whether they represented a genuine instance of an alternating verb. Because there was substantial agreement (κ = 0.705, Landis and Koch 1977; Viera and Garrett 2005: 262–263), the remaining combinations were judged by a single annotator, yielding 101 alternating verbs. Here, we focus on the 13 verbs alternating with the preposition *naar* 'to'.

Testing the hypotheses

First, a separate Belgian and Netherlandic regression model was composed based on the instances of all verbs combined, where theme argument was expressed and that had been manually marked as interchangeable. These models contained the fixed effects VERB (all different verbs), THEME COMPLEXITY (natural logarithm of the number of words of the theme), VERB-THEME ORDER (*theme-before-verb, verb-before-theme*), an interaction term between both, and a random effect with random intercepts for CORPUS COMPONENT. The categorical fixed effects were implemented through dummy coding. These regression models yielded the first effect plot on the front of this sheet. It was found that the lectal hypotheses were confirmed, as the Netherlandic data exhibit more outspoken lexical biases of the verbs than the Belgian data. Next, we investigated the alternation for individual verbs. Here, the results for *verlangen* 'desire' and *peilen* 'gauge' are shown. We ran collostructional analyses on the object slots of the transitive *verlang*-construction and the *verlang-naar*-construction to identify the 5 most prototypical slot fillers of both constructions. This was done separately for Belgium and the Netherlands. We then calculated distributional vectors for these 5 collexemes, as well as all other full nominal theme roots. The vectors used dependency-based context features of 8 possible relations as in Levshina and Heylen (2014: 30). Context features with function words or with the verb *verlangen* 'desire' itself were disregarded. Only the 5000 most frequent context features were used in the vectors, and the frequencies were weighted through positive point-wise mutual information. Finally, we calculated for each theme root t the measure COHERENCE TO THE *VERLANG-NAAR*-CONSTRUCTION as in the following equation (for sim_{cm}, see Weeds, Weir and McCarthy 2004). This was again done separately for Belgium and the Netherlands. Finally, a Belgian and Netherlandic regression model were built with the Belgian resp. Netherlandic measure as a fixed effect.

$$\text{COHERENCE TO THE } \textit{VERLANG-NAAR}\text{-CONSTRUCTION of theme root } t = 10\left(\frac{\sum_{n=1}^{5} sim_{cm}\left(\overrightarrow{naar-cxn\ collexeme_n}, \vec{t}\right)}{5} - \frac{\sum_{n=1}^{5} sim_{cm}\left(\overrightarrow{transitive-cxn\ collexeme_n}, \vec{t}\right)}{5}\right)$$

To control for higher-level semantic influence, we also used collostructional analyses to identify the 5 top collexemes of the object slots of the abstract transitive and *naar*-construction. We then calculated COHERENCE TO THE *NAAR*-CONSTRUCTION of each object root, analogously to the calculation above, and added this as a fixed effect to the regression models. Furthermore, the fixed effects THEME COMPLEXITY, VERB-THEME ORDER, and an interaction between both were added, as well as a random effect with random intercepts for OBJECT ROOT (all different object roots). To get the model to converge, all object roots that occurred only once in the data were binned into a rest category for the random effect OBJECT ROOT, and the random effect CORPUS COMPONENT had to be left out. Only instances with full nominal object roots were included from the data, as no value of the semantic coherence measures could be calculated for the other instances. The instances of the collexemes themselves were also excluded to avoid circularity. These models yielded the second and third effect plots on the front of this sheet. Finally, the same was done, mutadis mutandis, for the verb *peilen* 'gauge' based on the Belgian data. A Netherlandic model could not be composed, since the prepositional variant was nearly non-existent in the Netherlands. This time, the random effect with random intercepts CORPUS COMPONENT could be maintained in the model. This model yielded the fourth effect plot at the front of this page. The VIF's of all models were well below 5.

Acknowledgments

We cordially thank Kris Heylen for methodological advice, as well as Katrien Beuls, Paul Van Eecke and Melanie Röthlisberger for helpful comments on an earlier draft of this paper. We also owe thanks to Peter Petré, Timothy Colleman and Martin Hilpert for highly valued input, as well as to other participants of the ICCG-10 conference in Paris.

References

Goldberg, A. E. (1999). The emergence of the semantics of argument structure constructions. In B. Macwhinney (Ed.), *Emergence of Language* (pp. 197–212). Hillsdale: Lawrence Earlbaum Associates.
Gries, S. T. (2007). Coll.analysis 3.2a.
Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.
Levshina, N., & Heylen, K. (2014). A radically data-driven Construction Grammar: Experiments with Dutch causative constructions. In R. Boogaart, T. Colleman, & G. Rutten (Eds.), *Extending the Scope of Construction Grammar* (pp. 17–46). Berlin: Mouton de Gruyter.
Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In P. Spyns & J. Odijk (Eds.), Essential Speech and Language Technology for Dutch (pp. 219–247). Heidelberg: Springer.
Perek, F. (2015). Argument structure in usage-based construction grammar. Amsterdam: John Benjamins.
R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna.
Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. IJCL, 8(2), 209–244.
van Noord, G. (2006). At Last Parsing Is Now Operational. TALN, 20–42.
Viera, A., & Garrett, J. (2005). Understanding interobserver agreement: the kappa statistic. Family Medicine, 37(5), 360–363.
Weeds, J., Weir, D., & McCarthy, D. (2004). Characterising Measures of Lexical distributional similarity. In COLING '04: Proceedings of the 20th international conference on Computational Linguistics (p. 1015).

Pijpops, Dirk, Dirk Speelman, Stefan Grondelaers and Freek Van de Velde. 2018. Comparing explanations for the Complexity Principle. Evidence from argument realization. *Language and Cognition* 10(3). 514–543.