Constructional contamination: a pervasive effect

Dirk Pijpops, Isabeau De Smet & Freek Van de Velde

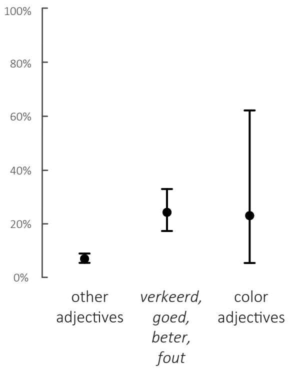
QLVL, University of Leuven
Research Foundation Flanders (FWO)

Constructional contamination is the effect whereby a subset of instances of a target construction is (stochastically) affected in its realization by a contaminating construction, due to a coincidental resemblance between the superficial strings of these instances and a number of instances of the contaminating construction.

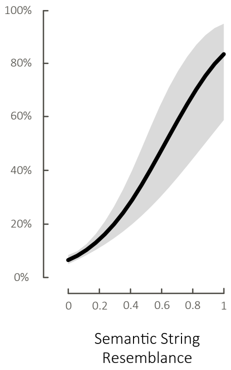
CASE STUDY 1: PARTITIVE GENITIVES

- Target: *ik heb iets verkeerd/iets verkeerds gegeten.*
  'I have eaten something wrong.'
- Contaminating: adverbs, *ik heb iets verkeerd geïnterpreteerd.*
  'I have wrongly interpreted something.'

Estimated probability of a partitive genitive without *-s*



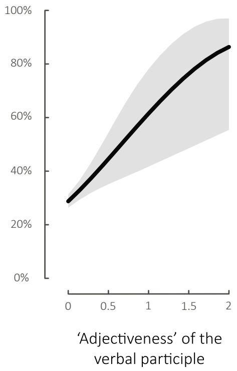
Estimated probability of a partitive genitive without *-s*



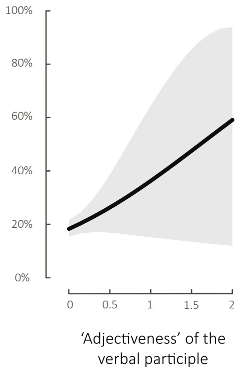
CASE STUDY 2: VERBAL CLUSTERS

- Target: *dat de deur door John gesloten is/is gesloten.*
  'that the door has been closed by John.'
- Contaminating: adj + copula, *dat de deur lange tijd gesloten is.*
  'that the door has been shut for a long time.'

Estimated probability of the participle + auxiliary order for auxiliaries *zijn* 'be' and *worden* 'become', which are also used as copulae



Estimated probability of the participle + auxiliary order for auxiliary *hebben* 'have', which is not used as a copula



CASE STUDY 3: WEAK VS. STRONG PRETERITES

- Target: *ik graafde/groef een put.*
  'I was digging a hole.'
- Contaminating: enclitic 2nd person, *waarom graafde een put?*
  'why are you digging a hole?'

CASE STUDY 4: LONG VS. BARE INFINITIVES

- Target: *Als ze de hele les zitten te slapen/*$^{?}$*zitten slapen.*
  'if they are sleeping throughout the entire class.'
- Contaminating: Infinitivus Pro Participio (IPP)*: ze hebben de hele les zitten slapen*, 'they have slept throughout the entire class.'

Theoretical importance

- Shallow parsing & storage of ready-mades
- Superficial similarities in usage affect grammar
- Horizontal links between constructions

Methodological importance

- Identify new case studies
- Find superficially resembling constructions
- Apply one of the available quantitative measures

Pijpops, Dirk and Freek Van de Velde. 2016. Constructional contamination: How does it work and how do we measure it? In: Folia Linguistica. 50(2): 543-581.

Pijpops, Dirk, Isabeau De Smet and Freek Van de Velde. Submitted. Constructional contamination in morphology and syntax: four case studies.

Materials & Methods

For the first case study, 3018 partitive genitives were extracted from the ConDiv-corpus (Grondelaers et al. 2000), of which 2700 were marked as strictly unambiguous, 2276 with *-s* ending and 424 without. Controlling for all factors known to influence *-s* omission as well as random lexical preferences, a strong effect of constructional contamination was found: the measure SEMANTIC STRING RESEMBLENCE correlated with a predilection for -s omission (p < 0.001, Odds Ratio = 4.36).

For the second case study, De Sutter provided a dataset containing 1440 unequivocal verbal clusters with a participle and the auxiliaries *zijn* 'be' or *worden* 'become', of which 1005 in the PARTICIPLE + AUXILIARY order, and 435 in the AUXILIARY + PARTICIPLE order, as well as 664 verbal clusters with a participle and the auxiliary *hebben* 'have', of which 126 in the PARTICIPLE + AUXILIARY order, and 538 in the AUXILIARY + PARTICIPLE order. The more often a verbal participle was used as an adjective in other sentences, the stronger it preferred the PARTICIPLE + AUXILIARY order. That is, we found the measure ADJECTIVENESS to correlate with a preference for this order among auxiliaries *zijn* 'be' and *worden* 'become' (p = 0.001, Odds Ratio = 3.96), and among the auxiliary *hebben* 'have', though not significantly (p = 0.132, Odds Ratio = 2.54).

For the third case study, 3641 instances of alternating verbs were extracted from a Twitter corpus compiled by Tom Ruette, yielding 3490 strong forms and 151 weak forms. Controlling for verb frequency and random lexical preferences, we found greater weakening in the regions where enclitic 2nd persons are part and parcel of the spoken dialects (p = 0.031, Odds Ratio = 0.395). This corroborates earlier findings of Vosters (2012: 242), that were based on elicited data.

For the fourth case study, 2766 instances of potential bare infinitives were extracted from the Sonar-corpus and manually checked (Oostdijk et al. 2013). In this way, we identified 7 bare infinitives where a present plural verb forms a cluster with an infinitive, thereby rendering the cluster superficially identical to a contaminating IPP-cluster. This contrasts with 2622 long infinitives in the same condition. We also found 3 bare infinitives where another finite verb and infinitive form a cluster that superficially resembles, yet is not identical to, a contaminating IPP-cluster. This contrasts with 11,978 long infinitives in the same condition. Finally, we detected 1 bare infinitive that was not part of a verbal cluster and was therefore not affected by constructional contamination from the IPP-construction. This contrasts with 13,576 long infinitives in the same condition. The differences in prevalence of bare infinitives between the first and second groups and between the first and third groups are both significant, with p < 0.001 (Fisher's exact test).

Acknowledgments

This research benefited from a BOF-grant at the University of Leuven and funding from the Research Foundation Flanders (FWO). We want to thank Gert De Sutter for generously sharing his richly annotated dataset of two-verb clusters collected as part of his PhD (De Sutter 2005), as well as Tom Ruette for giving us access to his Twitter-corpus.

References

Dąbrowska, Ewa. 2014. Recycling utterances: A speaker’s guide to sentence processing. *Cognitive Linguistics* 25(4). 617–653.

De Sutter, Gert. 2005. *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen*. Dissertation University of Leuven.

Ferreira, Fernanda & Nikole Patson. 2007. The “good enough” approach to language comprehension. *Language and Linguistics Compass* 1. 71–83.

Grondelaers, Stefan, Katrien Deygers, Hilde Van Aken, Vicky Van den Heede & Dirk Speelman. 2000. Het CONDIV-corpus geschreven Nederlands [The CONDIV-corpus of written Dutch]. *Nederlandse Taalkunde* 5(4). 356–363.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste & Ineke Schuurman. 2013. The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch. In Peter Spyns & Jan Odijk (eds.), *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, 219–247. Heidelberg: Springer.

Pijpops, Dirk & Freek Van de Velde. 2014. A multivariate analysis of the partitive genitive in Dutch. Bringing quantitative data into a theoretical discussion. *Corpus Linguistics and Linguistic Theory.* Published online, ahead of print.

Vosters, Rik. 2012. Geolinguistic data and the past tense debate. Linguistic and extralinguistic aspects of Dutch verb regularization. In Gunther De Vogelaer & Guido Seiler (eds.), The dialect laboratory. Dialects as a testing ground for theories of language change, 227–248. Amsterdam/Philadelphia: John Benjamins.