

Authors: Dirk Speelman, Dirk Geeraerts

University/Affiliation: University of Leuven - RU Quantitative lexicology and variational linguistics, University of Leuven - RU Quantitative lexicology and variational linguistics

Email addresses: Dirk.Speelman@arts.kuleuven.be, Dirk.Geeraerts@arts.kuleuven.be

Putting the (in)direct causation hypothesis to the test: a quantitative study of Dutch *doen* ‘make’ and *laten* ‘let’

In this paper we analyze the choice between the Dutch causative verbs *doen* ‘make’ and *laten* ‘let’ in patterns of the form NP CAUSE [NP V (...)] in which CAUSE is a form of either *doen* or *laten*, V is an arbitrary infinitive and (...) stands for zero or more constituents which complete the embedded clause. Examples taken from our dataset are given in (1) and (2).

- (1) *Ze hebben iemand anders met de caravan laten terugkomen.*
‘They have let someone else return with the camper.’
- (2) *Als je ze doet tennissen tegen hun zin dan gaan ze niet veel vooruitgang boeken.*
‘If you make them play tennis against their will then they will not make much progress.’

The dataset was restricted to cases where it is clear from the context that *doen* ‘make’ and *laten* ‘let’ express causation. Most notably, cases where *laten* ‘let’ expresses permission rather than causation were excluded from the dataset. For instance, in example (1) it was clear from the context that the sentence should be interpreted as ‘They have arranged for someone else to return with the camper’, and not as ‘They have given someone else permission to return with the camper’.

1. Theoretical starting-point

Our theoretical starting-point is the *(in)direct causation hypothesis* that was first formulated by Suzanne Kemmer and Arie Verhagen (Verhagen & Kemmer 1992, Kemmer & Verhagen 1994, Verhagen & Kemmer 1997, Verhagen 1998, Verhagen 2000) and that was more recently analyzed in depth in Ninke Stukker’s PhD thesis (Stukker 2005). Drawing on Talmy’s notion of force dynamics (Talmy 1988, 2000), the (in)direct causation hypothesis crucially involves the role of the causee in the causative event. The (in)direct causation hypothesis states that the choice for either *doen* or *laten* is influenced by the degree of involvement of the causee. In Stukker’s words, in the case of direct causation, as expressed by *doen*, “The causer produces the effected event directly; there is no intervening energy source ‘downstream’”. In the case of indirect causation, as expressed by *laten*, “Besides the causer, the causee is the most immediate source of energy in the effected event; the causee has some degree of ‘autonomy’ in the causal process” (Stukker 2005: 50). We will argue in the paper that starting from this assumption about the conceptual difference between *doen* and *laten*, the following more specific hypotheses may be formulated about the distribution of both verbs.

- 1) If *doen* expresses direct causation, we may expect more *doen* with animate matrix

- subjects: animate subjects have more control over the flow of energy.
- 2) If *laten* expresses indirect causation, you don't expect *laten* in constructions with an intransitive infinitive V: in the pattern NP CAUSE [NP V] the second NP typically is the ultimate affectee and the causee is not expressed.
 - 3) If *doen* expresses direct causation, coreferentiality between causer and causee or causer and affectee should favour the use of *doen*: you cannot get more direct than when you exert an influence on yourself.
 - 4) If the relevant factors are purely semantic ones, as in the (in)direct causation model, we don't expect any collocational idiomatization of the distribution: lexical fixation effects should not occur if the distribution is determined by conceptual factors only.
 - 5) At a conceptual level direct causation may be regarded to be the prototypical case of causation, so if *doen* expresses direct causation, its meaning is the center of the causative construction as a whole and can we expect those V infinitives which are themselves typically associated with causative constructions (because of their semantics) to favour *doen*.

2. Dataset and variables

The corpus we used for our case study is the Spoken Dutch Corpus (*CGN - Corpus Gesproken Nederlands*). The Spoken Dutch Corpus (see e.g. Oostdijk 2002 and Schuurman et al. 2003), compiled between 1998 and 2003, contains about 9 million tokens of contemporary spoken standard Dutch. It contains 14 different registers. From this corpus we collected 3975 occurrences of the pattern NP CAUSE [NP V (...)] and we encoded them for the following variables.

The variable *cause*, with possible values *doen* and *laten*, expresses the choice of causative verb and serves as the response variable in the statistical analysis which is discussed in the next section. The following predictors are used to test the specific hypotheses we derived from the general (in)direct causation hypothesis: the variable *inanim* stands for 'inanimateness of the first NP'. Its possible values are *no* and *yes*. The variable *cstr* stands for 'construction type'. Its possible values are *intransitive* and *transitive*, which stand for intransitive V and transitive V respectively. The variable *coref* stands for 'coreferentiality'. Its possible values are *no* and *yes*, which stand for complete absence of coreferentiality versus presence of some type of coreferentiality respectively. The variable *sig.lex.col* stands for 'significant lexical collocation' (at an alpha-level of 0.05), and it has two possible values: *yes* and *no*. The information we want to store in this variable pertains to 'lexical fixation'. We want to establish whether in some (or many) of the items in our dataset there is (some degree of) lexical fixation at play in the link between the infinitive V and the specific causal verb (either *doen* or *laten*). For this we use a method which is essentially a collocation analysis (Stefanowitsch & Gries, 2003) although we establish significance by means of the log likelihood ratio test which was introduced into linguistics by Dunning (1993). The variable *sig.sem.col* is designed to capture 'significant semantic (or conceptual) collocations', as opposed to the more conventional 'significant lexical collocations' captured by *sig.lex.col*. The variable *sig.sem.col* is designed to reflect whether there is a significant attraction between the infinitive at hand and the 'abstract causative construction as such' (making abstraction of the specific causative verb). The rationale behind the variable is that verbs which are attracted to the infinitive slot of causative constructions, do so because their meaning easily links up with the concept, i.e. the

semantics, of causation. This rather less conventional type of collocation analysis will be discussed at length in the paper.

Apart from the (in)direct causation hypothesis related variables we also added two variables by means of which we want to verify some additional variationist assumptions. The predictor country, with possible values nl (for The Netherlands) and be (for Belgium) simply encodes whether an observation is drawn from the Netherlandic Dutch or the Belgian Dutch part of the Spoken Dutch Corpus. The predictor spont, with possible values yes and no, simply encodes whether an observation is drawn from the spontaneous speech part (yes) or the prepared speech part (no) of the Spoken Dutch Corpus.

3. Logistic regression analysis

Table 1 lists results from the logistic regression analysis. Variable selection was obtained through forward as well as backward selection (the results were identical). The obtained statistical model is not a simple one since there are some interaction terms (which will be discussed in detail in the paper), but still the overall conclusion must be that several of the (in)direct causation hypothesis induced specific hypotheses were not confirmed by the data, most notably hypotheses 1), 3) and 4).

Table 1: predictor estimates and p values for the logistic regression model

| predictors (in order of introduction in forward stepwise regression) | estimates (positive is pro 'doen') and p- values for model with main effects and two way interactions |
|--|---|
| (intercept) | -3.26 (p < 0.001) |
| inanim (yes) | 3.57 (p < 0.001) |
| country (be) | 1.08 (p < 0.001) |
| sig.sem.col (yes) | 1.28 (p < 0.001) |
| sig.lex.col (yes) | 2.33 (p < 0.001) |
| sig.lex.col:sig.sem.col | -3.41 (p < 0.001) |
| cstr (transitive) | -0.36 (p = 0.25) |
| cstr:sig.sem.col | -1.50 (p < 0.001) |
| spont (yes) | -0.95 (p < 0.001) |
| coref (yes) | -1.23 (p = 0.006) |
| inanim:spont | 1.23 (p = 0.01) |
| cstr:spont | 0.67 (p = 0.047) |

4. Interpretation of results

We believe that the case study sheds new light on the (in)direct causation hypothesis. Although this study is no more than a first step towards a thorough quantitative test of that hypothesis, it nevertheless is a substantial one. Although the study does not imply that the hypothesis should be abandoned entirely, it does narrow down the number of

legitimate interpretations of the hypothesis. We will argue in the paper that we need to rethink and refine the (in)direct causation hypothesis on the basis of our findings. We will also suggest an alternative interpretation of the results, which approaches the functional differences between *doen* and *laten* from a different angle.

References

- Dunning, Ted 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61-74.
- Kemmer, Suzanne & Arie Verhagen 1994. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5, 115-156.
- Oostdijk, Nelleke 2002. The design of the Spoken Dutch Corpus. In: Pam Peters, Peter Collins and Adam Smith (eds.), *New Frontiers of Corpus Research*, 105-112. Amsterdam: Rodopi.
- Schuurman, Ineke, Machteld Schoupe, Heleen Hoekstra and Ton Van der Wouden 2003. CGN, an annotated corpus of spoken Dutch. In: Anne Abeillé, Silvia Hansen-Schirra and Hans Uszkoreit (eds.), *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora*, 101-108. Budapest, Hungary.
- Stefanowitsch, A. and Gries, S.T. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2:209-43.
- Stukker, Ninke 2005. Causality marking across levels of language structure. PhD dissertation, University of Utrecht.
- Talmy, Leonard 1988. Force dynamics in language and cognition. *Cognitive Science* 12: 49-100.
- Talmy, Leonard 2000. *Toward a cognitive semantics*. Cambridge: MIT Press.
- Verhagen, Arie & Suzanne Kemmer 1997. Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics* 27, 61-82.
- Verhagen, Arie 1998. Changes in the use of Dutch *doen* and the nature of semantic knowledge. In Ingrid Tieken-Boon van Ostade, Marijke van der Wal & Arjan van Leuvensteijn (eds.), *DO in English, Dutch and German. History and present-day variation*, 103-119. Amsterdam/Münster: Stichting Neerlandistiek/Nodus Publikationen.
- Verhagen, Arie 2000. Interpreting Usage: Construing the history of Dutch causal verbs. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-Based Models of Language*, 261-286. Stanford, CA: CSLI Publications.