Dutch causative constructions

Quantification of meaning and meaning of quantification

Natalia Levshina, Dirk Geeraerts, and Dirk Speelman
University of Marburg / University of Leuven

This chapter is a multivariate corpus-based study of two near-synonymous periphrastic causatives with *doen* and *laten* in Dutch. Using multiple logistic regression and classification trees, the study explores the conceptual differences between the constructions. The results support the existing definition of *doen* as the direct causation auxiliary, and interpretation of *laten* as the indirect causative (e.g. Verhagen and Kemmer 1997). However, the analyses also reveal more specific patterns: the most distinctive semantic pattern of *doen* is affective causation, whereas the contexts with the highest probability of *laten* refer to inducive causation. These differences remain valid when we control for geographic and thematic variation, as well as for the individual Effected Predicates treated as random effects in a mixed model.

Keywords: classification trees, logistic regression, mixed model, periphrastic causatives

Uncorrected proofs - © John Benjamins Publishing Company

1. Introduction

This chapter is a contribution to empirical Cognitive Semantics (e.g. Glynn and Fischer 2010).[1] It is a corpus-based multivariate onomasiological study (cf. Tummers *et al.* 2005), which uses quantitative corpus evidence to describe, explain and predict the choices that speakers make between semantically related constructions when they categorize their experience. To do so, the linguist needs to identify the relevant semantic, pragmatic, social and other features that influence this choice. This kind of

1. This research was supported with a grant from the Flemish Research Fund – FWO (G033008). The authors would also like to thank Kris Heylen for his help in collecting the corpus data. The usual disclaimers apply.

study requires advanced statistical multivariate techniques, such as logistic regression, which allow the researcher to model the impact of each factor, while controlling for the others.

The approach is relatively new, but a number of studies have been implemented already. It is interesting to note that a substantial share of these studies focus on constructions that differ from each other with respect to information structure and processing. Examples are the dative alternation in English (e.g. Bresnan *et al.* 2007), presence or absence of the presentative *er*-construction in Dutch (Grondelaers *et al.* 2007), particle placement in English (Gries 2003), word order variation in Dutch final verbal clusters (de Sutter 2009) and in the German 'middle field' (Heylen 2005). It seems that these alternations, a challenge for traditional linguistic descriptions, have benefited the most from the multifactorial probabilistic methods due to a variety of ways in which the underlying information-processing factors can be captured. However, more "semantic" constructional variation, like the one discussed here, can benefit from these methods too because highly abstract grammatical meaning can be captured in a corpus by a multitude of indirect indicators used as circumstantial evidence.

The current chapter, which is an elaboration of the pilot study carried out by Speelman and Geeraerts (2009), focuses on the near-synonymous Dutch causative constructions with *doen* and *laten*. Our study incorporates several linguistic, thematic and geographical factors in a multivariate statistical model, which allows us to test the existing semantic hypotheses about *doen* and *laten*, keeping the conceptual factors apart from the other sources of variation. We argue that the distinctive conceptual features that emerge in the quantitative model constitute the distinctive prototypes of the constructions – the semantic configurations with the highest intercategorial cue validity.

Although the Prototype Theory of categorization (Rosch 1975; Rosch and Mervis 1975) has been dominant in Cognitive Linguistics, many psychological and, more recently, linguistic studies (e.g. Medin and Schaffer 1978; Bybee and Eddington 2006) have demonstrated the crucial role of specific exemplars (or low-level schemata) in category organization and development. This is why we also test whether the abstract semantic differences between the constructions still hold if we take into account the lexemes that fill in the effected predicate slot, many of which display strong preference for *doen* or *laten*. The method applied for this purpose is mixed-effect modelling with the general semantic and other factors as fixed effects and the specific effected predicates as random effects.

The chapter has the following structure. First, we give a brief introduction of the Dutch causative constructions. In Section 3, the data and the potentially relevant variables are presented. Section 4 reports the results of the multiple logistic regression analysis and additional tests, which are interpreted linguistically and cognitively in Section 5. The chapter ends with a summary of our findings.

Uncorrected proofs © John Benjamins Publishing Company

2. Dutch causative constructions

Modern standard Dutch has two periphrastic causatives with the infinitive: the constructions with *doen* 'do' and *laten* 'let'. They share the same schematic pattern: an initiator causes another entity to acquire a state or perform an action. Consider example (1):

(1) *De politie deed/liet de auto stoppen.*
the police did/let the car stop
'The police stopped the car'.

Using the terminology from Kemmer and Verhagen (1994), *de politie* 'the police', is the causer of the event; *de auto* 'the car' is the causee that performs the action specified by the effected predicate *stoppen* 'stop'. The forms *deed* and *liet* are the past forms of the causative auxiliaries *doen* and *laten*, respectively. The most striking feature of the Dutch causatives is that *laten* as a causative auxiliary can refer both to the enabling and coercive types of causation (see Verhagen and Kemmer 1997: 69). Compare the situations in (2a), (2b) and (2c):

(2) a. *De trainer liet de spelers loopoefeningen doen.* [coercive]
the coach let the players running-exercises do
'The coach made the players do running exercises'.
b. *Hij liet iedereen zijn roman lezen.* [ambiguous]
He let everyone his novel read
'He made/had/let everyone read his novel'.
c. *De politie liet de dader ontsnappen.* [enabling]
the police let the criminal escape
'The police let the criminal escape'.

There have been a number of usage-based studies that have tried to establish the differences between the constructions (Kemmer and Verhagen 1994; Verhagen and Kemmer 1997; Degand 2001; Stukker 2005; Speelman and Geeraerts 2009). Verhagen and Kemmer (1997) write about the semantic difference between *doen* and *laten* in terms of the speaker's conceptualization of the situation as direct or indirect causation, respectively. Direct causation means that "there is no intervening energy source 'downstream' from the initiator: if the energy is put in, the effect is the inevitable result" (Verhagen and Kemmer 1997: 70). Indirect causation, which also includes the situations of enablement and permission, emerges when the situation "can be conceptualized in such a way that it is recognized that some other force besides the initiator is the most immediate source of energy in the effected event" (*Ibid.*: 67).

Speelman and Geeraerts (2009) showed, in their multivariate analysis of the Corpus of Spoken Dutch, that there is also a substantial amount of geographic and register variation in the use of the constructions. From the conceptual point of view, it

Uncorrected proof © John Benjamins Publishing Company

was suggested that *doen* is an obsolescent form with a tendency towards semantic and lexical specialization, most probably in direct physical causation (Speelman and Geeraerts 2009: 200), although this hypothesis was not tested.

The highly abstract conceptual patterns, such as direct and indirect causation, cannot be directly observed in a corpus-based study. Our aim is to explore a set of independent contextual factors ("diagnostic features", according to Speelman and Geeraerts 2009) that can serve as indirect, or circumstantial, evidence of semantic differences between the constructions. These contextual factors, operationalized as independent variables in the logistic regression model, are listed in Section 3, as well as the extralinguistic (geographic and thematic) variables that are explored in this study.

3. Data and variables

3.1 Data

The study is based on an 8 million token corpus of Netherlandic and Belgian Dutch, compiled from the TwNC and LeNC newspaper corpora (2001–2002). The corpus was balanced with regard to four subject domains of the articles: politics, economy, football and music. We used a syntactically parsed version of the data, which was obtained with the help of the Alpino parser of Dutch (Bouma *et al.* 2001). This allowed us to extract the contexts with constructions automatically. The contexts were then checked manually to avoid spurious hits and formally similar but functionally different constructions, such as the adhortative *laten* in *Laten we gaan* 'Let's go'. We also excluded idiomatic expressions with effected predicates that do not occur independently, e.g. *begaan*, which only occurs in the set expression *laten begaan* 'release, give freedom'. After the manual cleaning, we were left with 6,808 observations, which were then coded for seven semantic, syntactic, geographical and thematic variables presented in the next section.

3.2 The response variable

The speaker's choice for *doen* or *laten* in the given context was used as the binary response variable. The distribution of the constructions in the data set was skewed towards *laten*, which occurred 5,636 times, while *doen* was used only in 1,172 contexts, which is approximately 5 times less.

Uncorrected proofs - © John Benjamins Publishing Company

3.3 The linguistic predictors

The variable *CrSem* refers to the semantic class of the causer: animate (humans and animals) or inanimate (material and abstract entities). All previous studies reported the more frequent use of animate causers with *laten* and inanimate ones with *doen*. Verhagen and Kemmer (1997) and Stukker (2005) studied the causer's semantics only in combination with the semantic class of the causee. They found, however, that inanimate causers in combination with both animate or inanimate causees tend to be used more frequently with *doen* because these configurations correspond to physical and affective causation types, respectively, which imply direct causation. The most typical configuration for *laten*, which normally represents inducive causation, consists of animate causers and causees. This type of causation is indirect because humans cannot influence other humans' minds directly, telepathy disregarded (Verhagen and Kemmer 1997: 71). The remaining possibility, the combination of animate causers with inanimate causees, allows both for direct and indirect interference of the causer.

CeSem stands for the semantic class of the causee, which can also be animate or inanimate. If the causer is not the main source of energy in indirect causation, then it should most probably be the causee (cf. Stukker 2005). Thus, one could expect a higher degree of animacy of the causee in the *laten*-construction in comparison with *doen*. This variable has not been examined separately in any of the previous studies, although from Verhagen and Kemmer's (1997) description of causation types it follows that the chances for inanimate causees to be used with *doen* are somewhat higher than for animate ones. Both explicit and implicit causees (see below) were classified, depending on the context and the semantics of the effected predicate. Nevertheless, we were unable to classify 13 cases with implicit causees, so we left those contexts out.

CdEventSem describes the semantic class of the caused event. It can be mental or non-mental (physical or social). In case of metaphorical meaning, we assigned the semantic class that corresponded to the target domain. For example, in (3) the caused event was coded as mental:

(3) *Het doet het belletje rinkelen.*
it makes the bell_DIM ring
'It rings a bell'.

This variable was included to test whether *doen* is associated with the physical causation, as Speelman and Geeraerts (2009) suggested.

EPTrans refers to the transitivity (including ditransitivity) or intransitivity of the effected predicate. The previous studies showed that *laten* is more favoured by transitive verbs, which was regarded as evidence for the indirectness of the causative situations indicated by this construction because it involved a longer causation chain with more participants.

Uncorrected proofs - © John Benjamins Publishing Company

The variable *CeSynt* was inspired by Kemmer and Verhagen's (1994) observations about the *laten*-construction. In some contexts, the causee in Dutch allows not only for zero-marking as in (4a), but also for the prepositions *aan* and *door*, the dative and instrumental/agentive markers in Dutch, respectively, as in (4b) and (4c):[2]

(4) a. *Hij liet **zijn vrouw** zijn nieuwe gedicht lezen.*
'He made/let his wife read his new poem'.
b. *Hij liet zijn nieuwe gedicht **aan zijn vrouw** lezen.*
'He let his wife read his new poem.'
c. *Hij liet zijn nieuwe gedicht **door zijn vrouw** lezen.*
'He had his new poem read by his wife'.
d. *Hij liet zijn nieuwe gedicht lezen.*
'He had his new poem read'.

Kemmer and Verhagen (1994) argue on the basis of cross-linguistic evidence that propositional or indirect-object marking of the causee implies a smaller degree of integration of the causee into the causative event and its lower affectedness in comparison with the default zero-marking (or, for personal pronouns, marking with the case of the direct object). This smaller integration and affectedness is typical of indirect causation. Therefore, we should expect prepositional marking to boost *laten*.

On the other hand, Kemmer and Verhagen also suggested that implicitness of the causee, like in (4d), means even larger peripherality and non-affectedness of the causee (Kemmer and Verhagen 1994: 139). However, a more recent research study by Loewenthal (2003) has shown that implicit causees in the *laten*-construction have a moderate degree of affectedness, although the peripherality claim may still hold. Considering all this, and also the low frequencies of the prepositional marking, we distinguished two levels of the predictor: "Central" (the causee is explicit and unmarked) and "Peripheral" (the causee is implicit or marked with a preposition).

4. Statistical analysis

Multiple logistic regression allows us to model the speaker's behaviour by taking into account several factors that influence the speaker's choice simultaneously. The analyses were carried out with the help of R statistical software (R Development Core Team 2010). The first step of multivariate analysis is the selection of variables that have an impact on the speaker's choice. To select the relevant variables, we used the forward and backward stepwise selection procedures based on Akaike's Information

2. Note that all three marking options are available only for one effected predicate, *lezen* 'read'. The *aan*-marking is typical for verbs of perception and, consequently, mental causees, whereas the preposition *door* normally marks agentive causees.

Uncorrected proofs - © John Benjamins Publishing Company

Table 1. Results of multiple regression (simple main effect model)

Predictor	Estimate (log odds ratio)
(Intercept)	−4.38 ($p < 0.001$)
CrSem = Inanimate	3.44 ($p < 0.001$)
EPTrans = Intransitive	1.48 ($p < 0.001$)
Country = BE	0.68 ($p < 0.001$)
CdEventSem = Mental	0.79 ($p < 0.001$)
CeSynt = Peripheral	−0.90 ($p < 0.001$)
SubjectDomain = Football	0.12 ($p = 0.38$)
SubjectDomain = Music	0.45 ($p < 0.001$)
SubjectDomain = Politics	0.35 ($p = 0.009$)

Criterion (AIC). This criterion helps strike a balance between the predictive power of a model and its parsimony. All predictors, except *CeSem*, entered the final model (see Table 1).[3]

The order of the variables in the table reflects their importance in predicting the speaker's behaviour, as selected by the forward stepwise algorithm. The most important predictor is *CrSem*, and the least influential one is *SubjectDomain*. The column with the estimates provides the log odds ratios of the *doen*-construction for the given value of the predictor in comparison with the reference level (the values of the variables not mentioned in the table: the animate causer, the transitive effected predicate, the Netherlands, the explicit zero-marked causee, the non-mental effected predicate, and Economy as the article's subject domain). If the log odds ratio is equal to 0, *doen* and *laten* have equal chances to occur, which means that the predictor is not informative. A positive value means that the chances of *doen* are higher for the given value in comparison with the reference level of the same predictor. A negative log odds ratio, conversely, stands for relatively higher chances of *laten*. For example, the inanimate causer increases the log odds ratio of *doen* in comparison with the reference level, the animate causer, by 3.40, which corresponds to the simple odds ratio of 29.96 (i.e. the chances for inanimate causers to occur in the *doen*-construction are almost 30 times as high as those of the animate causers). The *p*-values next to the estimate demonstrate how confident one can be that the estimate is not equal to 0: the lower the *p*-value, the more certain one can be. Conventionally, a value of $\alpha = 0.05$ is used as a cut-off point for significant effects.

The overall quality of the model is satisfactory, as the measurements in the left-hand column in Table 3 suggest. The most intuitive measure is the proportion of correct predictions of *doen* and *laten* by the model. We can correctly predict 90.1% of

3. Not all statisticians agree on the value of stepwise selection (e.g. Harrell 2001). However, analysis of the full model and single term deletion tests yield the same model structure.

Uncorrected proof © John Benjamins Publishing Company

the speakers' choices (the cut-off probability is set to 0.5). However, this is not the most informative measure because *laten* is so frequent that if we simply predicted *laten* for all contexts, we would be correct in 82.8% of the cases, which serves as the baseline. There are special measures that neutralize the skewness, for example, the index of concordance *C*, also called the area under the ROC curve (see Hosmer and Lemeshow 2000: 160), which is 0.893 in our case. This number means that for all pairs of contexts with *doen* and *laten*, the model assigns a higher probability to the auxiliary actually observed in the context in almost 90% of the cases. *C* is equal to 0.5 if the predictions are random, and is equal to 1 if they are perfect. The related measures are Somers' D_{xy} = 0.787 (rank correlation between the predicted probabilities and observed responses ranging from 0 to 1) and Goodman-Kruskal's Gamma = 0.795 (with the range from –1 to 1). R^2 is Nagelkerke's generalized R^2 index for logistic regression models, which is analogous to the measure of explained variation in linear models. It stands for a proportional reduction in the absolute value of the log-likelihood measure in comparison with the intercept-only model (see e.g. Menard 2001: 24–27 for more details). It ranges from 0 (no predictive power) to 1 (a perfect fit of the data). For this model, R^2 = 0.523. All these values demonstrate that the model has a substantial predictive power.

However, one more thing should be taken into account. The effect of some of the predictors on the response variable may be non-additive, i.e. it cannot be explained by the summary effect of the predictors taken separately. An example of such an interaction from health care is a situation when one and the same medical treatment produces different effects on patients depending on their sex or age. In this study, we focused on the interactions between the intralinguistic variables (semantic and syntactic ones). We selected a model with interaction terms on the basis of AIC. The procedure was performed in such a way that all five semantic and syntactic variables in all possible combinations had a chance to occur in the model. One three-way interaction *CrSem:EPTrans:CeSynt* turned out to be significant. Table 2 displays the model with the interaction (it also lists the relevant lower-order interaction terms).

In a model with interaction terms, interpretation of the coefficients is less straightforward because we can no longer estimate an independent impact of a predictor that participates in an interaction without taking into account different levels of the other variables with which it interacts. For example, the estimate for *CrSem* = Inanimate in Table 2 should be interpreted as the combination of *CrSem* = Inanimate AND *EPTrans* = Transitive AND *CeSynt* = (the two latter terms are the reference levels of the corresponding variables).

Table 3 lists the summary statistics for the two models. One can see that the predictive power of the model with interactions is slightly better in comparison with the main-effect only model, although not dramatically. This shows that our main-effect only model was informative, but too coarse to deal with some combinations of the predictor values.

Uncorrected Proofs - © John Benjamins Publishing Company

Table 2. Model with main effects and three-way interactions

Predictor	Estimate (log odds ratio)
(Intercept)	−3.59 ($p < 0.001$)
CrSem = Inanimate (for *EPTrans* = Transitive and *CeSynt* = Central)	3.67 ($p < 0.001$)
EPTrans = Intransitive (for *CrSem* = Animate and *CeSynt* = Central)	0.41 ($p = 0.051$)
Country = BE	0.68 ($p < 0.001$)
CdEventSem = Mental	0.78 ($p < 0.001$)
CeSynt = Peripheral (for *EPTrans* = Transitive and *CrSem* = Animate)	−1.93 ($p < 0.001$)
SubjectDomain = Football	0.16 ($p = 0.27$)
SubjectDomain = Music	0.49 ($p < 0.001$)
SubjectDomain = Politics	0.34 ($p = 0.014$)
EPTrans = Intransitive: *CeSynt* = Peripheral (for *CrSem* = Animate)	3.48 ($p < 0.001$)
CrSem = Inanimate: *CeSynt* = Peripheral (for *EPTrans* = Transitive)	−0.31 ($p = 0.437$)
CrSem = Inanimate: *EPTrans* = Intransitive (for *CeSynt* = Central)	0.26 ($p = 0.459$)
CrSem = Inanimate: *EPTrans* = Intransitive: *CeSynt* = Peripheral	−2.60 ($p < 0.001$)

Table 3. Summary statistics for two models

Statistic	Model without interactions	Model with interactions
Number of observations	6795 (*doen*: 1170, *laten*: 5625)	
Proportion of correct predictions (baseline = 82.8%)	90.1%	90.3%
C	0.893	0.91
D_{xy}	0.787	0.821
Gamma	0.795	0.829
Generalized R^2	0.523	0.553
AIC	3695.8	3525.2

Three-way interactions are hard to grasp intuitively, so we used another technique, named CART (Classification And Regression Trees), to visualize the interactions in a convenient way. The algorithm splits up the observations according to the values of each predictor, trying to separate the observations with *doen* from those with *laten* in the best possible way. It begins with the split that allows the cleanest separation, and then proceeds with the resulting subsets, choosing the next best split.

Uncorrected proofs - © John Benjamins Publishing Company

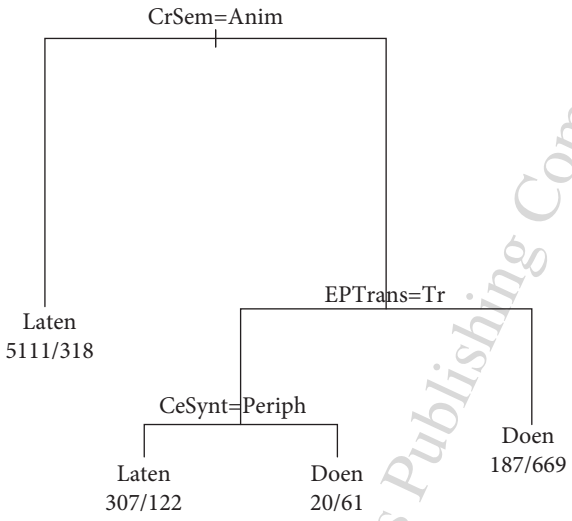

Figure 1. Classification tree for semantic and syntactic variables

The procedure then cross-validates the resulting tree against different subsets of data, selecting the most parsimonious model with the purest "leaves".

Figure 1 shows the classification tree for our dataset (only the semantic and syntactic variables took part in the classification). It was implemented with the help of the rpart package in R.[4] The minimum number of observations allowed in a split was 20; the algorithm performed 10 cross-validations. Under these conditions, only three of the five linguistic features take part in splits: *CrSem*, *EPTrans* and *CeSynt*. Recall that they are also the ones that interact significantly in the regression model. Each split is labeled with a decision rule, e.g. "CrSem=Anim". If the condition is met, i.e. for all animate causers, one should follow the left branch; otherwise, the right branch should be explored. The names of the leaves display the predominant construction in the group; the numbers below stand for the number of *laten*- and *doen*-observations. The error rate of the classification was low: 9.5%, in comparison with 17.2% if we simply always predicted *laten* as the default auxiliary (cf. the baseline in Table 4).

The first observation one can make is that *doen* needs more conditions to be met (inanimate causer AND intransitive effected predicate, or, in very few cases, inanimate causer AND transitive effected predicate AND explicit unmarked causee), whereas the animate causer is perfectly sufficient to obtain a leaf with a sufficient probability of *laten*, which also contains the largest number of observations. In addition, *laten*

4. An alternative solution is to use conditional inference trees. Their main advantage is that they neutralize the bias towards covariates with many possible splits (see Hothorn *et al.* 2006). However, all linguistic variables in the present analysis are binary, so this factor should not cause problems.

Uncorrected proof – © John Benjamins Publishing Company

emerges in more specific situations with the inanimate causer, transitive effected predicate and peripheral causee. The classification also tells us that the features *EPTrans* and *CeSynt* are relevant for the classification only in the case of the inanimate causer. For animate causers, these features are not powerful enough to influence the outcome. In a similar way, the syntactic expression of the causee has a decisive effect only in the case of an inanimate causer and a transitive effected predicate.

5. Linguistic interpretation of the statistical models

Some of our expectations based on the (in)direct causation hypothesis were confirmed by the marginal effects of the variables in the simple main effect model: inanimate causers, intransitive effected predicates and syntactically central causees do favour *doen*. However, we found no indication that the semantic class of the causee is significant, although one might expect that an animate causee is a better candidate for an indirect causation event because it is the main source of energy in the causation process. This lack of evidence creates a dilemma that is common in empirical studies (Geeraerts 1999). On the one hand, it may cast doubt on the indirect-direct causation hypothesis in the way it was formulated above. Alternatively, one could question our operationalization of the causee's role in terms of animacy or inanimacy. The latter scenario seems to be more reasonable. The fact that *doen* is preferred by mental caused events suggests that the causation categorized with *doen* frequently involves animate causees as experiencers. In contrast, the *laten*-construction, preferred by more dynamic non-mental caused events, contains animate causees, who can play a more active, agentive role. Therefore, other, more sophisticated ways of determining the causee's role could be helpful.

The observed preference of *doen* by mental caused events is unexpected. In combination with inanimate causers, it seems that *doen* is highly associated with affective causation, which involves a stimulus (causer) that triggers a mental reaction of an experiencer (causee). This behaviourist-like causation type is very direct. However, Verhagen and Kemmer's (1997) theory did not predict the predominance of affective causation within the semantics of *doen*; Speelman and Geeraerts (2009) even spoke about direct physical causation as *doen*'s specialization.

Next, we calculated the probabilities of *doen* and *laten* for every configuration of the linguistic features, as predicted by the model with interactions. To do so, we first calculated the sum of the relevant estimates provided in Table 2, including the intercept value, and then transformed the resulting log odds ratios into probabilities.[5] The

5. According to the formula $P = \exp(x)/(1+\exp(x))$, where x is the sum of the log odds ratios (coefficients) for all variables (the relevant values) and the intercept.

Uncorrected proofs © John Benjamins Publishing Company

configurations with the highest probabilities of *doen* and *laten* can be illustrated by contexts (5) and (6), respectively.

(5) *Het artikel (…) deed mij terugdenken aan mijn ontmoeting met de Algerijnse ambassadeur in Brussel, voorjaar 1972 op een receptie in Den Haag.*
'The article (…) made me think back to my meeting with the Algerian ambassador in Brussels in the spring of 1972 at a reception in The Hague'.

(6) *Prinses Juliana zou in de jaren 60 liefdesbrieven aan haar dochters hebben laten analyseren door een grafoloog.*
'They say that in the 1960s Princess Juliana had love letters to her daughters studied by a graphologist'.

Context (5), where the probability of *doen* is 89.7%, contains an inanimate causer, an intransitive effected predicate, an explicit unmarked causee and a mental caused event. This is a typical example of affective causation. Context (6) illustrates the configuration with the highest probability (99.2%) of *laten*, combining an animate causer, a transitive effected predicate, a peripheral causee and a non-mental caused event. The example evokes the service frame, when the causer uses the causee's professional services to have some work done. Note that both examples contain animate causees, who play different semantic roles (an experiencer and agent, respectively).

In his linear discriminant analysis of verb particle placement, Gries (2003) interprets the clusters of attributes with a highest distinctive load (highest discriminant scores of the sentences) as prototypes of each of the two constructions that he contrasts. Can we follow this approach and claim that contexts like (5) and (6) exemplify the prototypes of the constructions with *doen* and *laten*, respectively?

Indeed, it was shown by Rosch and Mervis (1975) that the features that help maximally distinguish the given category from the others are also the features that are maximally shared between the members of the same category. The presence of these features, operationalized as a family resemblance score of a category member, also correlated positively with the member's prototypicality in the category based on typicality ratings. This could lead us to the conclusion that the *doen*- and *laten*-observations such as (5) and (6), with the most distinctive features, should also be the best representatives of the categories from the intracategorial perspective.

However, the more recent studies of natural language categories (e.g. Ceulemans and Storms 2010) show that these kinds of salience do not always correlate. The results of additional experiments (Levshina 2011) show that significant positive correlations between the intra- and intercategorial types of salience operationalized in several different ways are observed only in the case of *doen*, and not in the case of *laten*. A possible explanation is that *laten*, which is used much more frequently than *doen*, is also a more heterogeneous category. In a polysemous category, the sense that is the most semantically distant from a contrasting category may not be central intracategorially. In addition, one can imagine that the distinctive features of *laten* with regard to

Uncorrected proof © John Benjamins Publishing Company

another construction may be different from those with regard to *doen*. All this means that we should be very specific about the perspective and operationalization when using the term 'prototype'. The configurations of the distinctive features exemplified by the sentences (5) and (6) are thus distinctive corpus-based 'prototypes' with regard to the choice between *doen* and *laten* modelled with the help of logistic regression.

In addition, the analyses reveal that *doen* is indeed quantitatively and semantically restricted, as Speelman and Geeraerts (2009) wrote. This can be illustrated by the classification tree in Figure 1, which shows clearly that a larger number of semantic and syntactic conditions should be satisfied for *doen* to have more chances to occur in a context than *laten*. Therefore, *doen* seems to have more Gestalt-like semantics than *laten*, which has a looser set of semantic features. This conclusion can be supported by the fact that *laten* is a highly schematic auxiliary with a semantic range from permission to coercion.

So far, we have not discussed the behaviour of the Subject Domain variable. The effect of the topic on the distribution of *doen* and *laten* was not predicted by any of the previous hypotheses. It would be natural to assume that different topics differ with regard to lexicon. This is why we looked at the top five most popular effected predicates in the four subject domains, which are shown in Table 4 with their relative frequencies.

One can see that the four topics are dominated by the same highly frequent verbs: *zien* 'see', *weten* 'know', *horen* 'hear', *denken* 'think', *liggen* 'lie', *vallen* 'fall', *gaan* 'go' with different relative frequencies. This might not be a serious problem if these highly frequent verbs did not demonstrate an outspoken preference either for *laten* or *doen*. For instance, *weten* is a typical *laten*-verb, with 450 occurrences in our data, all of them with *laten*. Some previous research, e.g. Levshina *et al.* (2009), which was based on collostructional analysis (Stefanowitsch and Gries 2003), also showed that attraction between the effected predicates and the constructions (auxiliaries) is indeed very strong. This is why the differences in the relative frequencies of these influential predicates may have an effect on the distribution of *doen* and *laten* in the subject domains. One of the ways to capture this idiomatic difference is to incorporate the lexical "noise" in the model as random effects. This method, called mixed-effect modelling, has proved to be a powerful tool in linguistic research, especially in

Table 4. Top five most frequent effected predicates in the four subject domains

Economy	Football	Music	Politics
zien 'see' 13%	*liggen* 'lie' 9%	*horen* 'hear' 11%	*weten* 'know' 10%
weten 'know' 11%	*zien* 'see' 8%	*denken* 'think' 8%	*zien* 'see' 6%
liggen 'lie' 5%	*weten* 'know' 5%	*zien* 'see' 4%	*vallen* 'fall' 4%
vallen 'fall' 3%	*vallen* 'fall' 4%	*klinken* 'sound' 3%	*gaan* 'go' 2%
stijgen 'go up' 3%	*spelen* 'play' 2%	*weten* 'know' 2%	*denken* 'think' 2%

Uncorrected proofs - © John Benjamins Publishing Company

psycholinguistic experiments with individual subject- and item-related noise (see a variety of case studies in Baayen 2008). It is also helpful in corpus-based studies when idiosyncrasies of individual words cannot be handled with the help of coarse-grained semantic classifications. Using the lmer package in R, we fit a mixed-effect with the effected predicates (1,165 types represented by 6,795 tokens) as random effects. By doing so, we "tell" the model that some effected predicates may inherently prefer *doen*, and that some verbs may prefer *laten*. The algorithm slightly lowers or increases the value of the intercept for each verb depending on its preferences in the data. It also takes into account the frequency of the verb in the data set. Ideally, we would have to do the same for the constructional slots of the causer and the causee. However, application of this method is less evident for nominal slots because of pronominal reference and lower type-token ratios. There is also evidence that the ties between nouns and constructions in which they appear are weaker than those between constructions and verbs (cf. Tomasello *et al.* 1997).

Fitting a mixed-effect model (with main effects only) yields the results shown in Table 5. The factors and the tendencies that we had in the corresponding model without random effects remain very similar, with the exception of the subject domain, which ceases to contribute substantially to the model's performance (according to AIC). Therefore, we can conclude that the difference in probabilities of *doen* and *laten* across different topics is due to the lexical effects. Also, most of the absolute values of the coefficients in the mixed model are slightly higher than those in the fixed-effect only model (see Table 1) because we have filtered out some part of the lexical "noise", which caused overfitting in the initial model.

The model demonstrates that the abstract features related to direct or indirect causation are still significant when conditioned on the lexical effects (cf. Bresnan *et al.* 2007: 87). At the same time, additional tests show that the random-effect model alone would allow the prediction of the choice of *doen* and *laten* correctly in a vast majority of cases (78% for the Netherlandic subcorpus and 74% for the Belgian data). This can be seen as evidence of strong exemplar effects at the level of the effected predicates.

Table 5. Logistic regression model with effected predicates as random effects

Predictor	Estimate (log odds ratio)
(Intercept)	–5.95 ($p < 0.001$)
CrSem = Inanimate	4.22 ($p < 0.001$)
EPTrans = Intransitive	2.25 ($p < 0.001$)
Country = BE	0.76 ($p < 0.001$)
CdEventSem = Mental	1.11 ($p < 0.001$)
CeSynt = Peripheral	–0.83 ($p = 0.003$)
SubjectDomain = Football	Not Available
SubjectDomain = Music	
SubjectDomain = Politics	

Uncorrected proofs - © John Benjamins Publishing Company

However, the best-performing model is the one with both the abstract and the lexical features. This finding is perfectly in line with the non-reductionist constructionist approach to language, which assumes that high-level generalizations coexist with low-level schemata in the speaker's knowledge about constructions (Langacker 1987; Goldberg 1995, 2006).

6. Conclusion

In this multivariate corpus-based onomasiological probabilistic study, we used logistic regression to find the factors that influence the choice between the causative constructions with *doen* and *laten* by speakers of Dutch. The analyses showed that the highest probability of *doen* is observed in the contexts of affective causation: an inanimate stimulus causing a conceptually and syntactically central cognizer to experience some mental state (an intransitive event). Conversely, the *laten*-construction has the highest chances of being observed when the causer is animate, the effected predicate is transitive, the causee is implicit or marked with a preposition (syntactically and conceptually peripheral), and the caused event is non-mental. This configuration can be exemplified by the service frame. The combinations of these features, which have the highest cue validity with regard to the choice between the categories, can be regarded as the distinctive prototypes of the constructions. Their relations with the other salience phenomena, such as family resemblance, goodness of membership, entrenchment, etc., are to be explored empirically, although there are indications that the inter- and intracategorial typicality measures tend to correlate for compact categories without rich polysemy (in our case, *doen*).

We also found evidence of exemplar effects in categorization at the level of the lexemes that fill in the effected predicate slot, which can serve as powerful predictors of the speaker's choice on their own. However, the best prediction is achieved when the model combines both the above-mentioned semantic generalizations and the lexemes. This supports the constructionist hypothesis that the mind stores linguistic knowledge at different levels of generalization.

In addition, the results show that the *doen*-construction has more chances of being chosen by a Flemish speaker than by a Dutch one, which supports the previous findings by Speelman and Geeraerts (2009) for spoken data and can be explained historically (the Flemish variety is believed to retain more archaic features). The constructions also display different behaviour across the four subject domains, which, as the mixed-effect model demonstrates, can be explained by the domain-specific differences in the distribution of the effected predicates.

Uncorrected Proofs - © John Benjamins Publishing Company

References

Baayen, R. H. (2008). *Analysing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511801686

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Boume, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.

Bouma, G., van Noord, G., & Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. In W. Dalemans, K. Sima'an, J. Veenstra, & J. Zavrel (Eds.), *Computational linguistics in the Netherlands 2000: Selected papers from the Eleventh CLIN meeting* (pp. 45–59). Amsterdam: Rodopi.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511750526

Bybee, J., & Eddington, D. (2006). A usage-based approach to Spanish verbs of 'becoming'. *Language*, 82(2), 323–355. DOI: 10.1353/lan.2006.0081

Ceulemans, E., & Storms, G. (2010). Detecting intra and inter categorical structure in semantic concepts using HICLAS. *Acta Psychologica*,133 (3), 296–304. DOI: 10.1016/j.actpsy.2009.11.011

De Sutter, G. (2009). Towards a multivariate model of grammar: The case of word order variation in Dutch clause final verb clusters. In A. Dufter, J. Fleischer, & G. Seiler (Eds.), *Describing and modeling variation in grammar* (pp. 225–254). Berlin & New York: Mouton de Gruyter.

Degand, L. (2001). *Form and function of causation. A theoretical and empirical investigation of causal constructions in Dutch*. Leuven: Peeters.

Geeraerts, D. (1999). Idealist and empiricist tendencies in cognitive semantics. In T. Janssen, & G. Redeker (Eds.), *Cognitive linguistics: Foundations, scope and methodology* (pp. 163–194). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110803464.163

Geeraerts, D. (2006). Salience phenomena in the lexicon: A typology. In D. Geeraerts (Ed.), *Words and other wonders: Papers on lexical and semantic topics* (pp. 74–97). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110219128.1.74

Glynn, D., & Fischer, K. (2010). *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*. Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110226423

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalizations in language*. Oxford: Oxford University Press.

Gries, St. Th. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York: Continuum.

Grondelaers, S., Geeraerts, D., & Speelman, D. (2007). A case for a cognitive corpus linguistics. In M. Gonzalez-Marquez, I. Mittleberg, S. Coulson, & M. Spivey (Eds.), *Methods in cognitive linguistics* (pp. 149–169). Amsterdam & Philadelphia: John Benjamins.

Harrell, F. E. (2001). *Regression modelling strategies with applications to linear models, logistic regression, and survival analysis*. Heidelberg & New York: Springer.

Heylen, K. (2005). A quantitative corpus study of German word order variation. In S. Kepser, & M. Reis (Eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives* (pp. 241–264). Berlin & New York: Mouton de Gruyter. DOI: 10.1515/9783110197549.241

Uncorrected Proof © John Benjamins Publishing Company

Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley. DOI: 10.1002/0471722146

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674. DOI: 10.1198/106186006X133933

Kemmer, S., & Verhagen, A. (1994). The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics*, 5(2), 115–156. DOI: 10.1515/cogl.1994.5.2.115

Langacker, R. W. (1987). *Foundations of cognitive grammar: Volume I: Theoretical prerequisites*. Stanford: Stanford University Press.

Levshina, N. (2011). *Doe wat je niet laten kan: A usage-based analysis of Dutch causative constructions [Do what you cannot let: A usage-based analysis of Dutch causative constructions]*. Unpublished doctoral dissertation, University of Leuven.

Levshina, N., Geeraerts, D., & Speelman, D. (2009). Collostructional analysis of Dutch causative constructions. *Paper presented at the Third International AFLiCo Conference*, 28 May, Paris.

Loewenthal, J. (2003). Meaning and use of causeeless causative constructions with *laten* in Dutch. In A. Verhagen, & J. van de Weijer (Eds.), *Usage-based approaches to Dutch* (pp. 97–130). Utrecht: LOT.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207–238. DOI: 10.1037/0033-295X.85.3.207

Menard, S. (2001). *Applied logistic regression analysis*. Thousand Oaks: Sage.

R Development Core Team (2010). *R: A language and environment for statistical computing. Foundation for statistical computing*. Vienna, Austria. <http://www.R-project.org>.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology*, 104(3), 192–233. DOI: 10.1037/0096-3445.104.3.192

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573–605. DOI: 10.1016/0010-0285(75)90024-9

Speelman, D., & Geeraerts, D. (2009). Causes for causatives: The case of Dutch *doen* and *laten*. In T. Sanders, & E. Sweetser (Eds.), *Causal categories in discourse and cognition* (pp. 173–204). Berlin & New York: Mouton de Gruyter.

Stefanowitsch, A., & Gries, St. Th. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–243. DOI: 10.1075/ijcl.8.2.03ste

Stukker, N. (2005). *Causality marking across levels of language structure*. Unpublished PhD dissertation, University of Utrecht.

Tomasello, M., Akhtar, N., Dodson, K., & Rekau, L. (1997). Differential productivity in young children's use of nouns and verbs. *Journal of Child Language*, 24(2), 373–87. DOI: 10.1017/S0305000997003085

Tummers, J., Heylen, K., & Geeraerts, D. (2005). Usage-based approaches in cognitive linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*, 1(2), 225–261. DOI: 10.1515/cllt.2005.1.2.225

Verhagen, A., & Kemmer, S. (1997). Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics*, 27(1), 61–82. DOI: 10.1016/S0378-2166(96)00003-3

Uncorrected proofs - © John Benjamins Publishing Company

Uncorrected proofs - © John Benjamins Publishing Company