Dirk Speelman & Dirk Geeraerts

Causes for causatives: the case of Dutch *doen* and *laten*

Throughout its recent history, mainstream linguistics has aspired to the uncontroversial legitimacy of the hard sciences: the Chomskyan statement that linguistics is a branch of theoretical biology is a claim to status just as much as it is a demarcation of a domain of enquiry. However, it is only in the last decades that linguistics has seriously begun to follow the most directly obvious way towards substantiating that claim: that of applying the scientific method. That is still far from being the dominant approach: in the context of Cognitive Linguistics for instance (the framework with which we are most familiar), meta-theoretical pleas for adopting the scientific method (Geeraerts 2006, Gibbs 2007, Gries 2006) contrast with nuanced but straightforward defenses of an introspective method, like Talmy (2007).

In the study that we are presenting here, we will illustrate the importance of the scientific method for linguistics by applying it to the description of *doen* and *laten* causatives in contemporary Dutch: given the existence of *doen* and *laten* as causative verbs, what is it that determines the choice between the two? The steps to be taken according to the principles of the scientific method will be obvious: from an initial theory about the phenomenon at hand we derive a set of predictions than can be tested against a sample of observable behavior, and that test might possibly lead to a falsification of the theory. So what are the elements that constitute the design of our investigation?

1. Background and research questions

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 flow of energy in the causative event. Some terminological clarification may be necessary at this point. In a pattern of the type NP1 CAUSE [NP2 V NP3], as illustrated by *the professor made the students follow the scientific method*, NP1 is the *causer*: the subject of the matrix sentence that is the most direct causal instigator of the event. NP3 is the *affectee*: the object of the embedded sentence that is the ultimately affected entity. And NP2 is the *causee*: the subject of the embedded sentence that functions as an intermediary between the causer and the affectee.

The (in)direct causation hypothesis now 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 ef-

fected 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). Starting from this assumption about the conceptual difference between *doen* and *laten* causatives, we may derive a number of hypotheses about the distribution of both verbs.

> Prediction 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.
>
> Prediction 2. If *laten* expresses indirect causation, you don't expect *laten* in constructions of the type NP1 CAUSE [NP2 V], as illustrated by *the professor made the students laugh*, where the embedded sentence is intransitive and where NP3 is not expressed: the causee, i.e. the intransitive subject of the embedded sentence, is the ultimate affected entity. (Constructions with an elliptical or pseudo-intransitive embedded sentence, of the type *the professor made the students submit* are counted as transitive.)
>
> Prediction 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 as when you exert an influence on yourself.
>
> Prediction 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.
>
> Prediction 5. At a conceptual level, direct causation may be regarded to be the prototypical case of causation, so if *doen* expresses direct causation, we expect those infinitives which are themselves typically associated with causative constructions (because of their semantics) to favour *doen*. If *doen* is associated with the (direct) core type of causation, we expect more *doen* in typically causative contexts.

Next to these predictions that may be derived from the (in)direct causation hypothesis, there is another piece of existing research that we need to take into account. Traditional descriptive work on the differences between Netherlandic Dutch and Belgian Dutch (that is to say, the national varieties of Dutch as used in The Netherlands and the Flemish part of Belgium respectively) point out that *doen* has a higher frequency in Belgian Dutch (see the lexicographical description in De Clerck 1981, Den Boon & Geeraerts 2005). Most of the variables that are known to mark a difference between Netherlandic Dutch and Belgian Dutch exhibit additional differences of register (Geeraerts, Grondelaers & Speelman 1999): more typically Belgian forms are found more often in informal (Belgian) registers. We may therefore formulate the following additional prediction.

> Prediction 6. The distribution of *doen* and *laten* is sensitive to lectal differences: we expect relatively more *doen* in Belgian sources than in Netherlandic sources, and within the Belgian sources, we expect more *doen* in in-

formal registers than in formal ones.

Two additional points need to be made with regard to this set of predictions. To begin with, the set is at the same time broader and narrower than the set of factors studied by Stukker (2005). Collocational effects and syntactic patterning (the distinction between transitive and intransitive embedded sentences, reflexive constructions) play only a minor role in her investigation, and since it is deliberately restricted to Netherlandic Dutch, lectal variation is not envisaged. Also, the quantitative analysis she presents is technically less advanced than the method we will use in the following pages: precisely because we incorporate more factors into the description, we need a method of analysis that is able to cope with such a complex set of data (this is a point we will come back to in a moment).

Conversely, Stukker focuses on the interplay between the animacy of the causer and the causee, while we have so far only incorporated the animacy of the causer. Note that we try to translate the hypothesis formulated by Stukker from a largely syntactic point of view, by focusing on the observable complexity of the argument structure of the constructions: is the 'interveninig energy source' explicitly expressed or not? An alternative approach would be to take a purely semantic perspective, and try to determine the 'autonomy' of the causee on independent grounds.

There is an important consequence of our choice of perspective that we need to make explicit: we consider the results to be presented here as a preliminary exploration of the field, which will need to be complemented with an even more extended scrutiny of potentially relevant variables. A research project carrying out this program has actually been started within our research group; this type of research is part of a broader line of research in which we develop a multivariate usage-based grammar. (For a representative sample of studies, see De Sutter, Speelman & Geeraerts 2005; Glynn, in press; Heylen 2005; Tummers, Speelman & Geeraerts 2005. For a theoretical statement, see Tummers, Heylen & Geeraerts 2005).

2. Limitations of the present study

Elaborating on the second point jst mentioned, we need to be explicit about the fact that the interpretation of the (in)direct causation hypothesis as we are testing it here, in the form of predictions 1-5, does not correspond to the interpretation presented by Kemmer, Verhagen, and Stukker. We are, in a sense, taking their formulation of the (in)direct causation hypothesis at face value, whereas their interpretation is semantically more subtle and complex. Their interpretation, in fact, associates (in)directness fairly directly with a distinction between physical causation and human causation, and hence, with the presence of animate or inanimate causers and causees.

Indirect causation refers to 'a situation that is 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' (Verhagen & Kemmer 1997: 6). Typically,

in a sentence like *De agent liet de studenten passeren* 'The police man let the students pass', the students ultimately do the passing: the policeman only creates the conditions for the students to perform the action. Following d'Andrade (1987), Kemmer and Verhagen further assume that we conceptualize situations with animate causers and causees in such a way that animate beings are not normally thought of as acting directly upon other human beings. While physical entities and forces are taken to exert a direct action on other things, animate beings exert an influence on others only indirectly, through the intervening medium of the physical world. It follows that direct causation is considered typical for physical, inanimate contexts, whereas animate contexts prime for indirect causation.

Given the Kemmer, Verhagen, and Stukker interpretation of the (in)direct causation hypothesis, why don't we take it directly as our point of departure? One reason we have already given: we explicitly see the present study as a first step, in which we illustrate the method on the basis of an initial operationalization, but which is leading up to a broader project in which we will consider alternative interpretations (including, needless to say, the original Verhagen and Kemmer one). Another reason for not starting straightaway from the Kemmer, Verhagen, and Stukker interpretation is of a methodological nature: we find the interpretation more difficult to operationalize, because it does not explicitly indicate how (in)direct causation should be detected independently.

To see the problem, let us have a closer look at the idea that causative contexts with an animate causer and an animate causee do not normally allow for direct causation. This assumption explains the predominant occurrence of *laten* in such contexts on the basis of the following implicit syllogism.

> First premise: A situation with an animate causer and an animate causee patterns with indirect causation.
> Second premise: Indirect causation patterns with *laten.*
> Conclusion: A situation with an animate causer and an animate causee patterns with *laten*.

Now, Kemmer and Verhagen present empirical evidence that *doen* is indeed more typical for inanimate, physical causation, and our own findings will not contradict that observation. However, does that suffice to conclude that direct causation patterns with *doen*, and indirect causation with *laten*? That would obviously only be the case if we had an independent diagnostic for establishing (in)directness of causation, and more particularly, if we could independently establish the first premise of the syllogism. As a simple point of logic, if our basic goal is to establish the second premise, we need independent evidence for the conclusion *and* for the first premise of the syllogism.

In practice, that would mean being able to determine whether sentences containing the causative construction exhibit direct or indirect causation regardless of whether the causative verb is filled out by *doen* or *laten.* How, in a sentence like *The teacher CAUSE the students finish their book*, could (in)directness be established? Kemmer, Verhagen and Stukker are not explicit about the criteria they would want to apply here, but such criteria would probably be fairly complex, se-

mantically speaking. For instance, while the description provided by Kemmer and Verhagen suggests that the distinction between physical and immaterial causation plays a role, a further analysis would have to refine that distinction and make it operational. Difficult cases are likely to occur: if *the lightning made the alarm go off* is material causation, should we then also classify *the lightning made the children tremble* as physical? The latter sentence probably features a less material type of causation than the former, but at the same time, *the lightning made the children tremble* would seem to be more material than *the idea of having to stay alone at home made the children tremble*. So where would we draw the line? The methodological point, however, would not be to enforce a binary categorial decision in every possible case, but rather to find an operationally applicable set of diagnostic features that would make it possible to chart all possible borderline cases and nuances. Such a componential analysis of the relevant contexts of use would indeed almost inevitably imply that the concepts 'direct' and 'indirect causation' stop being categorial variables, but rather reveal themselves as prototypical reference points on a continuum (or perhaps even in a multidimensional semantic space).

Elaborating a set of criteria for such a componential analysis is definitely a requirement for the broader study that we announced, but for the present exploratory purposes, we've opted for a more straightforward interpretation of (in)directness: we assume that the 'intermediate energy source' is syntactically represented by the causee, and explore a number of configurations in which this intermediate energy source would be more or less prominent.

3. The materials in the case study

The data in the case study were taken from 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, called the ‘components’ of the corpus, labelled A through N. They are listed in Table 1. The first column in Table 1 contains the label of the component. The second contains a short description. The other columns indicate which components contain dialogues or multilogues (DIA/MUL) and which contain monologues (MONO), which are spoken in a private context (PRIV) and which are spoken in a public context (PUB), and, finally, which contain spontaneous speech (SPON) and which contain more or less prepared speech (PREP). For each of the 14 components, the corpus contains data spoken by speakers from The Netherlands (henceforth Netherlandic Dutch) as well as data spoken by speakers from Belgium (henceforth Belgian Dutch). The exception is component E, for which there are no Belgian data. On average the amount of Netherlandic Dutch in the corpus is about twice as large as the amount of Belgian Dutch.

We automatically collected all instantiations in the corpus of the schematic pattern NP CAUSE [NP V (...)], in which CAUSE is a form of either *doen* or *laten*, V is an arbitrary infinite and (...) stand for zero or more constituents which complete

the embedded clause. After the initial automatic collection step we manually corrected the results in order to remove spurious hits. However, because of the nature of the annotation schema of the Spoken Dutch Corpus, which for most of the data in the corpus is restricted to lemmatization and part of speech tagging, we were forced to impose a restriction on our automatic data collection procedure; we restricted ourselves to those sentences in which there is either no or at most one token in between the form of *doen* or *laten* and the infinitive. In other words, our dataset does include sentences like *Ik wil je laten aanvoelen dat er een verschil is* 'I want to make you feel that there is a difference' and like *Hij liet me aanvoelen dat er een verschil is* 'He make me feel that there is a difference', but not *Hij liet de verbaasde menigte aanvoelen dat er een verschil is* 'He made the astonished crowd feel that there is a difference'. Not imposing this artificial restriction would have made the manual correction step prohibitively labor intensive. However, we are well aware of the artifical nature of the restriction and of the fact that its possible consequences for the results of this study will need further inspection in future research.

Table 1: the components of the Spoken Dutch Corpus

A	Spontaneous conversations ('face-to-face')	DIA/MUL	PRIV	SPONT
B	Interviews with teachers of Dutch	DIA/MUL	PRIV	SPONT
C	Spontaneous telephone dialogues (recorded via a switchboard)	DIA/MUL	PRIV	SPONT
D	Spontaneous telephone dialogues (recorded on MD with local interface)	DIA/MUL	PRIV	SPONT
E	Simulated business negotations	DIA/MUL	PRIV	SPONT
F	Interviews/ discussions/debates (broadcast)	DIA/MUL	PUB	PREP
G	(political) Discussions/debates/ meetings (non-broadcast)	DIA/MUL	PUB	SPONT
H	Lessons recorded in the classroom	DIA/MUL	PUB	SPONT
I	Live (e.g. sports) commentaries (broadcast)	MONO	PUB	SPONT
J	Newsreports/reportages (broadcast)	MONO	PUB	PREP
K	News (broadcast)	MONO	PUB	PREP
L	Commentaries/columns/reviews (broadcast)	MONO	PUB	PREP
M	Ceremonious speeches/sermons	MONO	PUB	PREP
N	Lectures/seminars	MONO	PUB	PREP

The manual correction had a double purpose. On the one hand we excluded a few

straightforward spurious hits such as *Dat moet te doen zijn* 'This must be feasible' which are no instantiations at all of the general pattern we look for. On the other hand we also excluded cases which can be considered instantiations of the construction, but which we chose to exlude either because of their special syntactic status or because of the impossibity of variation (at the synchronic level). One category we excluded are nominalizations such as *het laten varen van all hoop* 'Letting go of all hope'. Other categories we excluded are verbs that do no pattern independently such as *iemand laten betijen* 'to let someone be', optatives such as *laat ons hopen* 'let's hope', and grammaticalized idiomatic expressions such as *laat ons zeggen* 'let's say' or *laat staan dat* 'let alone that'.

4 The variables

In total 3975 observations survived the manual correction step. Having retrieved the data, we annotated them for the following variables.

4.1 The response variable `cause`

In section 4 we will present an analysis of our data in which we statistically model the choice that language users make (either consciously or inconsciously) for either the causal verb *doen* or the causal verb *laten* as a function of a series of factors such as the animacy of NP1, the transitivity of V, the presence of coreference, etc. What the statistical model will 'express', is whether these factors, the so-called predictors, indeed affect the probability of the chosen causal verb being one specific verb (e.g. *doen*). The choice of causal verb will be the so-called response variable in our statistical model: the variable the values of which we want to 'predict' with our model. We call our response variable `cause`. It has two possible values, `laten` and `doen`. This variable was encoded automatically, which obviously was a trivial procedure. In our dataset of 3975 observations we have 3664 cases of `cause=laten` and 311 cases of `cause=doen`. Clearly this is a heavily biased distribution with a proportion of 0.9218 (cases of `cause=laten`) versus a proportion of 0.0782 (cases of `cause=doen`).

4.2 The predictor `inanim`

The variable `inanim` stands for 'inanimateness of NP1'. Its possible values are `no` and `yes`, which stand for animate NP1 and inanimate NP1 respectively. This variable was encoded manually. Besides humans, animals as well as human collectives (*het team* 'the team', *de regering* 'the government', *de natie* 'the nation') were encoded as animate. In our dataset of 3975 observations we have 3776 cases of `inanim=no` and 199 cases of `inanim=yes`.

The purpose of this variable in our study is to test prediction 1 from section 1; if prediction 1 is accurate, then we expect the 'predictor state' `inanim=yes` to disfavour the response situation `cause=doen`.

4.3 The predictor `cstr`

The variable `cstr` stands for 'construction type'. Its possible values are `intransitive` and `transitive`, which stand for intransitive V and transitive V respectively. This variable too was encoded manually. In our dataset of 3975 observations we have 2124 cases of `cstr=transitive` and 1851 cases of `cstr=intransitive`.

The purpose of this variable in our study is to test prediction 2 from section 1; if prediction 2 is accurate, then we expect the 'predictor state' `cstr=transitive` to disfavour the response situation `cause=doen`.

4.4 The predictor `coref`

The variable `coref` stands for 'coreferentiality'. Its possible values are `no` and `yes`, which stand for complete absence of coreferentility versus presence of some type of coreferentiality respectively. This variable too was encoded manually. The following table gives a more explicit overview of the types of coreferentiality that are present in the dataset, and of the way we cope with them. In our dataset of 3975 observations we have 3654 cases of `coref=no` and 321 cases of `coref=yes`.

Table 2: the types of coreferentiality in the dataset

pattern	encoding `coref=no`	encoding `coref=yes`
x CAUSE ysubj Vintransitive	*ik CAUSE iets vallen* 'I CAUSE something fall'	*ik CAUSE mij vallen* 'I CAUSE myself fall'
x CAUSE ysubj Vtransitive	*ik CAUSE hem doen* 'I CAUSE him do'	*ik CAUSE mij doen* 'I CAUSE myself do'
x CAUSE zobj Vtransitive	*ik CAUSE iets zien* 'I CAUSE see something'	*ik CAUSE mij verrassen* 'I CAUSE myself be surprised'
x CAUSE ysubj zobj Vtransitive	*ik CAUSE iets iemand zien* 'I CAUSE someone see something'	*ik CAUSE iemand mij verrassen* 'I CAUSE someone surprise something'
x CAUSE zsubj *door* 'by' ypp Vtransitive	*ik CAUSE de deur door hem openen* 'I CAUSE the door be opened by him'	*ik CAUSE mij door iemand verrassen* 'I CAUSE myself be surprised by him'

The purpose of this variable in our study is to test prediction 3 from section 1; if prediction 3 is accurate, then we expect the 'predictor state' `coref=yes` to favour the response situation `cause=doen`.

4.5 The predictor `sig.lex.col`

The variable `sig.lex.col` requires a more lengthy explanation. The name of the variable `sig.lex.col` stands for 'significant lexical collocation', 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 causal verb CAUSE. For instance, if we encounter the sentence *Ik wil je iets laten weten* 'I want to let you know something' we want to establish whether there is (some degree of) lexical fixation between the infinitive *weten* 'know' and the causal verb *laten* 'let' and whether this fixation can be held responsible (at least to some extent) for the choice for *laten* 'let'. Informally speaking, we want to verify if *weten* 'know' triggers the choice for *laten* 'let', not (only) for semantic reasons such as the ones mentioned in the (in)direct causation hypothesis but (also) simply because the words *weten* 'know' and *laten* 'let' like to go together in causal patterns.

We operationalize lexical fixation on the basis of 'statistical collocation patterns'. Broadly defined, we speak of a statistical collocation pattern between a *word* and a *context* if the word and the context co-occur more often that would be expected on the basis of chance alone. Establishing statistical collocation patterns is done by means of a procedure called *collocational analysis*. Table 3 illustrates the general schema on which the concept of statistical collocation pattern and the procedure of collocational analysis are based.

Collocational analysis is always based on four frequencies. In Table 3 these frequencies are labelled `a`, `b`, `c` and `d`. Frequency `a` stands for the number of occurrences of the word under scrutiny in the context under scrutiny. For instance, if we want to establish whether there is a significant statistical collocation pattern between the word *weten* 'know' and the context 'infinitive V in a causal pattern with causal verb *laten*', then `a` stands for the number of times we encounter *weten* 'know' as the 'infinitive V in a causal pattern with causal verb *laten*'. The frequency `c` stands for the number of occurrences of an exhaustive range of other words in the context under scrutiny. In our example `c` stands for the number of times we encounter another word than *weten* 'know' as the 'infinitive V in a causal pattern with causal verb *laten*'. Having obtained these two frequencies we can use the ratio `a / (a+c)` as a measure for the popularity of *weten* 'know' in the context 'infinitive V in a causal pattern with causal verb *laten*'. The ratio has a straightforward interpretation: in `a` out of `(a+c)` cases the infinitive in the causal *laten*-pattern is *weten* 'know'.

Whereas the left column in Table 3 (the column with `a` and `c`) contains information

which is specific to the context under scrutiny, the right column (the column with `b` and `d`) serves as an external reference point. We need such a reference point because if in our example the ratio `a / (a+c)` is high, we are not yet sure that this is because *weten* 'know' prefers this context. An alternative explanation might be that *weten* 'know' is a high frequency word throughout the corpus, not just in the context under scrutiny. The ratio `b / (b+d)` can help us out, because it is a measure for the popularity of the word under scrutiny in a wide range of other contexts than the context under scrutiny. The frequency `b` is the sum of all occurrences of the word under scrutiny in any other context (out of a wide range of possible contexts) than the one we're interested in. The frequency `d` is the sum of all occurrences of any other word (in a wide range of possible words) than the word under scrutiny in the aforementioned wide range of other contexts than the one we're interested in. In our example the ratio `b / (b+d)` is a measure for the popularity of *weten* 'know' in a range of other contexts than 'infinitive V in a causal pattern with causal verb *laten*'. Once again, this ratio has a straighforward interpretation: in `b` out of `(b+d)` cases the word encountered in this range of other contexts is *weten* 'know'.

Table 3: general schema for collocational analysis

	in the context under scrutiny	in an exhaustive range of other contexts
number of occurrences of word under scrutiny	a	b
number of occurrences of an exhaustive range of other words	c	d

Now if `a / (a+c)` is higher than `b / (b+d)`, then we have detected a positive attraction between the *word* under scrutiny and the *context* under scrutiny, relative to the point of reference which `b / (b+d)` provides. We can subsequently use a statistical test to establish whether this attraction is statistically significant. Several statistical tests can be used. We will use the log likelihood ratio test which was introduce into linguistics by Dunning (1993).

For our implementation of the variable `sig.lex.col` we apply the schema in Table 3 along the lines of the approach we've been discussion on the basis of the example sentence *Ik wil je iets laten weten* 'I want to let you know something'. For each observation, i.e. each item in our dataset, we look at the actually used infinitive and the actually used causal verb, and we calculate a measure for the attraction between the two on the basis of the appropriate frequency information `a`, `b`, `c` and `d`. More precisely, we perform a collocational analysis in which the word under scrutiny is the observed infinitive and the context under scrutiny is the context 'infinitive in the causal construction with causal verb as observed'. If we can establish a significant attraction (at an alpha-level of 0.05) between the infinitive

and the causal verb, `sig.lex.col` receives the value `yes`, otherwise it receives the value `no`.

For each observation the frequency infomation `a`, `b`, `c` and `d` is derived from the complete Spoken Dutch Corpus as a whole. For instance, if the item in our dataset is the sentence *Ik wil je iets laten weten* 'I want to let you know something', the word under scrutiny is *weten* 'know' and the context under scrutiny is the context 'infinitive in the causal construction with causal verb *laten*'. The range of other words consists of all other words that occur in the Spoken Dutch Corpus and the range of other contexts consists of all occurrences of verbs (not just infinitives) in the Spoken Dutch Corpus in other positions than the infinitive position of the causal construction with causal verb *laten*'. In other words, in the example *ik wil je iets laten weten* we calculate:

a = *weten* in the context of causative *laten*
b = any other word than *weten* in the context of causative *laten*
c = *weten* in the context of other verbs than causative *laten*
d = any other word than *weten* in the context of other verbs than causative *laten.*

On a more technical note, we add that all counts are lemma based. This simply means that all different word forms of the same lemma are counted as instances of the same word.

Before we can conclude this lengthy discussion of `sig.lex.col` we must mention one additional rather technical issue. Although we just said that for all observations the calculations of `a`, `b`, `c` and `d` are based on the complete Spoken Dutch Corpus, this is in fact not true. We chose to calculate the lexical fixations captured by `sig.lex.col` differently for Belgian Dutch observations and Netherlandic Dutch observations, basing the calculations for `a`, `b`, `c` and `d` on the Belgian Dutch part of the Spoken Dutch Corpus in the case of Belgian Dutch observations and basing the calculations for `a`, `b`, `c` and `d` on the Netherlandic Dutch part of the Spoken Dutch Corpus in the case of Netherlandic Dutch observations. We proceeded in this way because we did not want to exclude the possibility that there exist (subtle) differences in lexical fixations between the two national varieties of Dutch.

In our dataset of 3975 observations we have 3051 cases of `sig.lex.col=yes` and 924 cases of `sig.lex.col=no`. Obviously these frequencies indicate that this type of procedure for establishing fixation (or rather attraction) patterns is calibrated differently from what a human researcher would consider fixation or no fixation. This automated procedure is sensitive to more subtle levels of fixation (or rather attraction).

We conclude the description of the variable `sig.lex.col` by defining the purpose of this variable. Its function is to test prediction 4 from section 1; if prediction 4 is accurate, then we expect the 'predictor state' `sig.lex.col=yes` to have no effect on the probability for the response situation `cause=doen`. In other words, the probability for the response situation `cause=doen` is expected to be the same in the

cases `sig.lex.col=yes` and `sig.lex.col=no`.

In technical terms, however, the test hypothesis will be that lexical fixation does have an effect on the preference for either of the causative verbs – only, we have no way of predicting the direction of the preference (either in favour of *doen* or in favour of *laten*). Starting from the idea that the (in)direct causation hypothesis is the only factor involved in the choice of the auxiliary, we expect that the test hypothesis will be disconfirmed. If that is indeed the case, we will have to be careful with the interpretation of the result: we will not exactly have found a confirmation of the (in)direct causation hypothesis, but we will at least not have observed a phenomenon that questions the hypothesis. Conversely, if the test hypothesis is confirmed, we will want to conclude that other factors besides the (in)direct causation hypothesis need to be taken into account to explain the behaviour of *doen* and *laten*.

4.6 The predictor `sig.sem.col`

Now we move to the variable `sig.sem.col`. This variable is designed to capture 'significant semantic (or conceptual) collocations', as opposed to the more conventional 'significant lexical collocations' captured by `sig.lex.col`.

The collocation analysis schema in Table 3 is a very flexible generic schema which can be applied to many different situations. Depending on the type of contexts one considers, the generic schema can be applied to very different types of analys: we can look for the keywords of a text, text type or register (contexts are texts, text types or registers), the typical neighbours of a target word (contexts are windows of words around a target word), the typical translations of a target word (context are translations of text fragments wich contain a target word), etc. The schema can easily be made even more generic by replacing the concept 'word under scrutiny' with the more generic 'item under scutiny', thus allowing for items which are smaller of larger than words.

In short, the schema can be applied in many different ways and indeed has been applied succesfully in many different ways. Also, the technique goes by many different names. In technical terms, the specific application on which our calculation of `sig.lex.col` is based is the same as the one underlying the procedure that is known in Cognitive Linguistics by the name of collostructional analysis (Stefanowitsch & Gries, 2003). The purpose of a collostructional analysis and our calculation of `sig.lex.col` is slightly different, though. In a typical collostructional analysis, you would determine the top-ranking infinitives that pattern significantly with a causative auxiliary like *laten*, and then determine the specific meaning of *laten* causatives on the basis of the semantic properties of those top-ranking verbs. In our case, by contrast, we use the degree of attraction between *laten* and such verbs (whether top-ranking or not) as a predictor variable in a statistical analysis.

In this section, we will go beyond collocational analysis in a technical sense as well, by introducing `sig.sem.col` as a relatively new way of applying the technique. In collostructional analysis one calculates the attraction between on the

one hand a lexical item such as *weten* 'know' and on the other hand a specific position (or 'slot') in a construction such as the V-slot in the construction NP1 *laten* [NP2 V (...)]. Typical of collostructional analysis is the fact that the construction under scrutiny has a lexically specific head (in this case *laten*). What we will do in this section, is perform a similar analysis for more abstract constructions, with a lexically unspecified head. In the example this would mean that we calculate the attraction between on the one hand a lexical item such as *weten* 'know' and on the other hand a specific position (or 'slot') in a more schematic construction such as the V-slot in the construction NP1 CAUSE [NP2 V (...)]. This abstract construction subsumes the cases in which CAUSE is *laten* and the cases in which CAUSE is *doen*.

The variable `sig.sem.col`, in other words, introduces *schematicity* into the analysis, in the sense in which it is known in Cognitive Linguistics (see Tuggy 2007). The variable is designed to reflect whether there is a significant attraction between the infinitive at hand and the 'abstract causative construction as such'. 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. Apart from the fact that we now have a larger context under scrutiny than in section 4.5, and apart from the fact that here we do not use separate calculations for Belgian Dutch and Netherlandic Dutch, the calculations for `sig.sem.col` are identical to those in section 4.5. Here too the possible values for the variable are `yes` and `no`, indicating presence and absence of a significant attraction respectively (at an alpha-level of 0.05). (We assume that there will be no major differences in the relevant conceptual preferences between Belgian Dutch and Netherlandic Dutch, but this is certainly an assumption that should be tested in further research.)

In our dataset of 3975 observations we have 2969 cases of `sig.sem.col=yes` and 1006 cases of `sig.sem.col=no`. Obviously, as for `sig.lex.col`, we must add that this automated procedure is sensitive to rather subtle levels of attraction.

The purpose of this variable in our study is to test prediction 5 from section 1; if prediction 5 is accurate, then we expect the 'predictor state' `sig.sem.col=yes` to disfavour the response situation `cause=doen`.

4.7 The predictors `country` and `spont`

The final two predictors will be introduced together. 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 (cf. Table 1).

In our dataset of 3975 observations we have 2395 cases of `country=nl` and 1580 cases of `country=be` and we have 2416 cases of `spont=yes` and 1559 cases of

spont=no.

Both variables are included in our study to test prediction 6 from section 1; if prediction 6 is accurate, then we expect the 'predictor state' country=be to favour the response situation cause=doen and we expect 'predictor state' spont=yes to also favour the response situation cause=doen (although a nuance may be that we might only expect the latter effect to be important if country=be).

4.8 Summary of the variables and their predicted effect

Having introduced all variables which we want to include in the statistical analysis in the section 4, we now present a summary table all the predictions we want to verify by means of the statistical analysis.

Table 4. Overview of the predictions we will test in the statistical analysis

id	predictor condition	predicted effect (test hypothesis)
predictions based on the (in)direct causation hypothesis		
1	inanim=yes	favours cause=doen
2	cstr=tr	disfavours cause=doen
3	coref=yes	favours cause=doen
4	sig.lex.col=yes	either favours or disfavours cause=doen
5	sig.sem.col=yes	favours cause=doen
predictions based on previous variationist research		
6	country=be	favours cause=doen
	spont=yes (especially when country=be)	favours cause=doen

5. The results of the statistical analysis

Logistic regression analysis is a type of regression analysis which is particularly suited for the situation in which the response variable has only two possible outcomes, such as in our case, where the possible outcomes are cause=laten and cause=doen. For a description of this technique and an introduction to its use in linguistics we refer to the specialized literature (e.g. Rietveld & Van Hout 1993: 327-361). In this text we will almost completely skip the technicalities and we will try to present the results from the analysis in such a way that the text is access-

ible to readers who are not familiar with the technique. For their convenience, we first give a very basic introduction to the interpretation of the regression output in section 5.1. Readers familiar with logistic regression analysis can safely skip section 5.1.

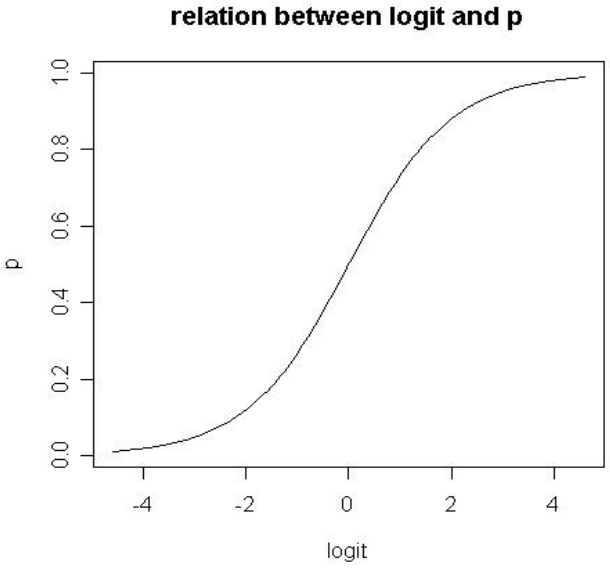
5.1 Reading the regression output

Put simply, the rationale behind regression analysis is that we describe either the sole effect of one predictor or, more typically, the combined effect of a series of predictors on the value of the response variable by means of a (typically rather simple) mathematical equation. This mathematical equation is called a 'statistical model' of the data. The equation predicts the values of the response variable in function of the values of the predictors. In practice these simple mathematical equations never completely accurately descibe reality - there typically are deviations -, but if the deviations from the model pattern like modest random noise we accept the model as a useful simplified representation of a more complex reality. In that case the interpretation of the (simple) mathematical equation can turn out to be very insightful.

The most attractive feature of the regression analysis procedure is that it is very capable (much more capable than the human researcher) to not just look at the effect of one predictor at the time, but specifically to look at the combined effect of several predictors, thereby identifying for each predictor what the effect of that predictor is 'when controlling for all other predictors'. This formulation means that the technique is good at seeing which difference one predictor makes in case 'all other predictors are held constant'.

Before we look at the results of our analysis, we have to address one technical aspect which is typical of logistic regression analysis (as opposed to other types of regression analysis) and which we believe is important for the interpretation of the results. Informally speaking, we can say a logistic regression model predicts the response value of an observation by assigning a 'probability p of having a specific response value' (in our analysis the model assigns a probability of 'cause=doen') to the observation by means of a mathematical equation. As in other types of regression analysis, this equation uses only the values of the predictors in that same observation to make the prediction. And obviously, a good model will on average assign a high probability for cause=doen to observations which actually have cause=doen and assign a low probability for cause=doen to observations which actually have cause=laten. A good model, in other words, is one in which the events that actually occur are recognized as highly probable, given the properties of the observations as measured by the predictor variables.

Figure 1: the relation between logit values and p values



However, for technical reasons, the logistic regression model does not actually assign a probability `p` of having e.g. `cause=doen` but rather a derived value, called `logit`, which is directly related to `p` but which is nevertheless a bit different. The `logit` is equal to `log(p/(1-p))` and conversely `p` is equal to `exp(logit)/(1+exp(logit))`. Fortunately this technical complication, however important, only modestly complicates the interpretation of the regression output, because the relation between `p` and `logit` is such that as the `logit` goes up, `p` goes up, and as the `logit` goes down, `p` goes down. We can see this in Figure 1, which plots the relation between `logit` and `p`. So we can remember that each time our statistical model predicts higher `logits`, we can infer that it also predicts higher probabilities for `cause=doen` and that each time it predicts lower `logits` we can infer from that that it also predicts lower probabilities for `cause=doen`.

We will present two logistic regression analyses of the data. Both models are represented in Table 5. For our first model, which we label 'model with main effects only', we ignore the right column in Table 5 and we also ignore the rows in Table 5 which are gray in the left column and which are empty in the middle column.

We obtained this first model by running several variable selection techniques (specifically forward stepwise regression and backward stepwise regression) in order to automatically select those variables (from the list of variables in Table 4 in section 4.8) which have a significant effect on the response variable and which make a significant contribution to the quality of the overall model. In these procedures, variables which fail the test are removed from the model. However, in this study we didn't have to leave out any variables from the model, since all variables turn out to have a significant effect on the response variable and on the overall quality of the model.

Table 5 can be read as follows: first of all, the order of the predictors reflects the

order in which the forward stepwise regression procedure selected the variables for inclusion in the model. This order is informative with respect to the relative importance of the predictors because the forward selection procedure first picks those variables which best reduce the amount of 'unexplained variation' in the model (i.e. the size of the deviations we mentioned before) thus improving the overall quality of the model. According to this criterion, `inanim` is the most important predictor, followed by `country`, followed by `sig.sem.col`, etc.

Next, we read the estimates of the model. Estimates reflect the effect of a predictor on the response variable, when controlling for other variables. We start with a special case, which is not really linked to a particular predictor: the estimate -2.73 for the so-called `(intercept)`, in the second row, is the `logit` which the model assigns to observations which for all predictors have the value which is *not* listed in the left column of Table 5. In other words, this is the `logit` the model predicts for observations which have `inanim=no`, `country=nl`, `sig.sem.col=no`, `sig.lex.col=no`, `cstr=intransitive`, `spont=no` and `coref=no`. A quick glance at Figure 1 tells us that a `logit` of -2.73 corresponds to a small probability. More precisely, it corresponds to a probability of 0.0612. So the category of observations just described are assigned a probability of 0.0612 for `cause=doen`, which is small but then again is not that extremely small if we recall that in the whole dataset the proportion of `cause=doen` is 0.0782. The probability assigned to the category of observations just described is only a bit below this global proportion.

From this point onwards we will call the predictor values *not* listed in the left column of Table 5 the *baseline* values of these predictors, and we will call the category of observations in which all predictors have their baseline value the baseline category of observations. (Note that apart from certain technical considerations the assigment of baseline status to a particular predictor value in principle is an arbitrary choice; in principle we could just as well have chosen other baseline values).

The estimates for the other predictors, i.e. 3.96 for `inanim`, 1.17 for `country`, -2.01 for `sig.sem.col`, etc. express a difference in predicted `logit` when the variable at hand does have the value listed in the left column of Table 5, as opposed to when it has the baseline value, and while 'controlling for all other predictors'. To give an example, all other things being equal (i.e. when controlling for other predictor variables), the model predicts the `logit` to be 3.96 higher if `inanim=yes`, when compared to the baseline case of `inanim=no`. And all other things being equal, the model predicts the `logit` to be 1.17 higher if `country=be`, as opposed to the case of `country=nl`. And all other things being equal, the model predicts the `logit` to be 2.01 lower if `sig.sem.col=yes`, as opposed to the case of `sig.sem.col=no`. And so forth. As was already mentioned, it is rather unfortunate from an interpretative point of view that these increases and decreases are expressed in `logits` and not in probabilities. Regrettably, there is no straightforward way to remedy this. When expressed on a `logit` scale, the effect of changing the value of one predictor is constant, irrespective of the values of the other predictors. However, if expressed on a probability scale, this effect would be variable, depending on the values of

the other predictors. In other words, when expressed as a difference in `logit` values, the offset in the prediction which is caused by a specific predictor is constant and does not depend on the values of the other predictors. But when expressed as a difference in probabilities, the offset in the prediction which is caused by a specific predictor depends on the probability one starts off with. This is because the relation in Figure 1 is S-curved and not linear.

Although we cannot simply express the effect of predictors on a probability scale, we can easily calculate the predicted probability of `cause=doen` for each different category of observations, as we already did for the baseline category `inanim=no`, `country=nl`, `sig.sem.col=no`, `sig.lex.col=no`, `cstr=intransitive`, `spont=no` and `coref=no`. For this baseline category the predicted `logit` is the estimate of the so-called `(intercept)`. In order to calculate the predicted `logit` for other categories of observations, we simply have to add to this `(intercept)` estimate the estimates for all predictors which do not have a baseline value in this category of observations. The simplicity of the mathematical equation resides in the fact that the effects of the different predictors simply have to be added up. For instance, for observations with `inanim=yes`, `country=be`, `sig.sem.col=no`, `sig.lex.col=yes`, `cstr=intransitive`, `spont=no` and `coref=no`, the predicted `logit` is -2.73 + 3.96 + 1.17 + 1.44, which is 3.87 and which corresponds (following Figure 1) to a predicted probability of `cause=doen` of 0.978. This is, incidentally, the highest probability of `cause=doen` assigned to any category of observations in this model. And to give another example, for observations with `inanim=no`, `country=nl`, `sig.sem.col=yes`, `sig.lex.col=no`, `cstr=transitive`, `spont=yes` and `coref=yes`, the predicted `logit` is -2.73 - 2.01 - 0.81 - 0.60 - 1.12, which is -7.27 and which corresponds to a predicted probability of `cause=doen` of 0.0006. This is the lowest probability of `cause=doen` assigned to any category of observations in this model.

There is one final piece of information in Table 5 which needs to explained: the p-values next to each estimate tell us how certain we are about the actual existence of an effect of that predictor on the response variable. If this value is below 0.05 we are (more than) 95% certain that the actual effect of this predictor differs from zero. In this case we call the effect of that predictor significant.

5.2 Main effects and two-way interactions

Before we discuss the results of the regression analysis, we briefly summarize what wat said in section 5.1. Table 5 shows the results from two separate regression analyses. Both models predict the logit for `cause=doen`. We start by discussing the model we see in the middle column. This model, which we label the 'model with main effects only', is the result of a stepwise forward regression as well as a stepwise backward regression procedure. Both procedures result in the same model in a trivial way: all variables are retained in the model. Table 5 is constructed is such a way that the order of the predictors reflects the order in which they were added to the model in the forward selection procedure.

Table 5: Estimates for two logistic regression models for the *doen-laten* data

predictors (in order of introduction in forward stepwise regression)	estimates (positive is pro 'doen') and p-values for model with main effects only		estimates (positive is pro 'doen') and p-values for model with main effects and two-way interactions	
`(intercept)`	-2.73	(p < 0.001)	-3.26	(p < 0.001)
`inanim (yes)`	3.96	(p < 0.001)	3.57	(p < 0.001)
`country (be)`	1.17	(p < 0.001)	1.08	(p < 0.001)
`sig.sem.col (yes)`	-2.01	(p < 0.001)	1.28	(p < 0.001)
`sig.lex.col (yes)`	1.44	(p < 0.001)	2.33	(p < 0.001)
`sig.lex.col:sig.sem.col`			-3.41	(p < 0.001)
`cstr (transitive)`	-0.81	(p < 0.001)	-0.36	(p = 0.25)
`cstr:sig.sem.col`			-1.50	(p < 0.001)
`spont (yes)`	-0.60	(p < 0.001)	-0.95	(p < 0.001)
`coref (yes)`	-1.12	(p = 0.01)	-1.23	(p = 0.006)
`inanim:spont`			1.23	(p = 0.01)
`cstr:spont`			0.67	(p = 0.047)

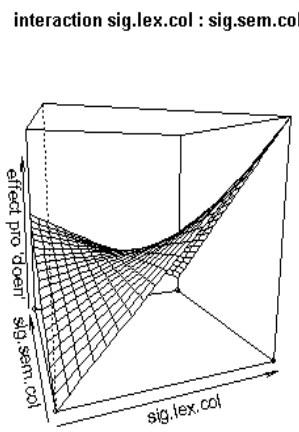
The 'model with main effects only' in the middle column of Table 5 is merely a stepping stone to the more complicated 'model with main effects and two-way interactions' in the right column. The 'model with main effects only' is presented because of its simplicity, but, as it turns out, it unjustly oversimplifies the patterns in the data. Further inspection of the data reveals that there are important two-way interactions.

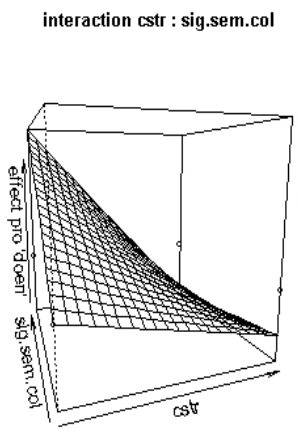
Therefore we introduce a second model. This model, labeled 'model with main effects and two-way interactions', is presented in the right column. Here too the order of the predictors reflects the order in which they were added to the model by the forward stepwise regression procedures. The new model contains the same predictors as the first model, which are so-called main effects, plus four new predictors, which are so-called interaction terms (or product terms). For instance, the presence of the interaction term `sig.lex.col:sig.sem.col` in the new model indicates that there is an interaction between `sig.lex.col` and `sig.sem.col` and that, in other words, the effect of `sig.lex.col` on the `logit` depends on the value for `sig.sem.col`, and conversely, the effect of `sig.sem.col` on the `logit` depends on the value for `sig.lex.col`. The joint effect of `sig.lex.col` and `sig.sem.col` cannot be expressed with two 'main effect' estimates only: the 'main effects only model' oversimplified matters at this point. Instead, it is the interplay of three estimates (`sig.lex.col`, `sig.sex.col` and `sig.lex.col:sig.sem.col`) which gives a more accurate account of this joint effect.

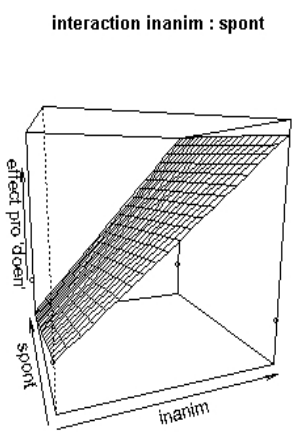
Because it is hard to interpret the joint effect of two interacting predictors on the basis of an output such as the one in Table 5 alone, we add visual representations of these joint effects. In Figure 2 we graphically represent the four interactions in our model. In these plots the x and y axes represent the interacting predictors and the z axis (the height) represents the joint effect of the two predictors on the `logit`. On the x and y axis the arrows points from the baseline values of the predictors (not listed in Table 5) to the alternative values (listed in Table 5). Three further remarks need to be made about the z axis. First, the plots are artificial in the sense that our predictors can assume only two possible values and that the only situations that can actually occur are represented by the four corners of the surfaces in the plot. All intermediate z-values are merely added to make the perspective of these three-dimensional representations easier to recognize. Second, although in the plots the z axis is represented on a `logit` scale, we will describe the effects in terms of increased or decreased predicted probability of `cause=doen` (remember that this probability goes up as the `logit` goes up and goes down as the `logit` goes down). Third, four small dots in the corners of each plot indicate the zero position on the y axis. This helps us to see whether joint effects are positive or negative.

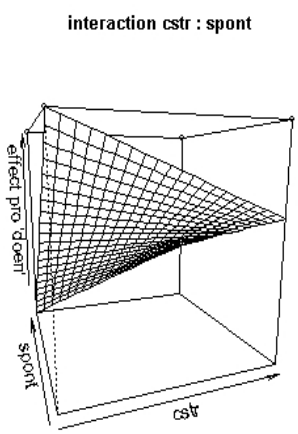
The most 'complicated' interaction is the one between `sig.lex.col` and `sig.sex.col`. The interaction is 'complicated' in the sense that a change of value in one predictor systematically reverses the effect of the other predictor. The complicated picture which emerges is that, compared to the baseline of [`sig.lex.col=no` and `sig.sem.col=no`] the effect pro `cause=doen` is slightly positive (`logit=0.2`) in case of [`sig.lex.col=yes` and `sig.sem.col=yes`], more positive (`logit=1.28`) in case of [`sig.lex.col=no` and `sig.sem.col=yes`] and most positive (`logit=2.33`) in case of [`sig.lex.col=yes` and `sig.sem.col=no`]. The other interactions, which are less complicated, but still important, can be interpreted in a similar way. We will not go over them step by step. We only draw attention to one other difference between the four plots. Whereas in the top left plot all effects are positive (`logits` range from 0.2 to 2.33), they are all negative in the bottom left plot (`logits` range from -0.36 to -0.95), and mixed in the top right plot (`logits` range from -0.36 to -1.28) and in the bottom left plot (`logits` range from -0.95 to 3.85).

Figure 2: A visual representation of the interactions in the second model









As a final piece of information regarding the analyses, we list a number of summary statistics for both analyses. We will not discuss these numbers, but do offer them so that the reader who is familiar with logistic regression analysis is more informed about the overall quality of the models.

Table 6: Summary statistics for the two logistic regression models

summary statistic	model with main effects only	model with main effects and two-way interactions
number of observations	3975 (of which 311 ‘doen’ and 3664 ‘laten’)	
null deviance	2181.9 (on 3974 df)	

residual deviance	1327.0 (on 3967 df) [AIC is 1343.0]	1228.2 (on 3963 df) [AIC is 1252.2]
model chi squared	854.8 (on 7 df)	953.7 (on 11 df)
p-value (chi squared)	$p < 0.001$	$p < 0.001$
simple proportion of correct predictions in original dataset (cut-off probability set to 0.5)	0.9494 (baseline is 0.9218)	0.9514 (baseline is 0.9218)
generalized R squared	0.4581	0.5050
C (area under ROC curve)	0.8754	0.9048

5.3 Interpretation of the results

When we now compare the results of the case study to the predictions which were formulated in the first section of the paper, we can draw the following conclusions. (The reader may corroborate the conclusions by comparing with the data in Table 5.)

> Prediction 1 is not confirmed in our dataset. Instead of favouring *doen* animate matrix subjects turn out to disfavour *doen* in all circumstances (in spite of the small interaction we found between animacy and the distinction spontaneous vs. prepared speech; this interaction does not reverse the effect of animacy).
>
> Prediction 2 is confirmed in our dataset, even if the picture is a bit more complicated than the prediction suggests. Transitive infinitives indeed disfavour *doen*, as was predicted, but they do so in a rather complicated way. Technically speaking: in the model with interactions the main effect of `cstr` is not significant but the variable is involved in two significant interactions (with `spont` and `sig.sem.col`). Still, further inspection of these interactions (top right and bottom right plots in Figure 2) shows that *doen* is disfavoured whenever the infinitive is transitive.
>
> Prediction 3 is not confirmed in our dataset. Instead of favouring *doen* coreferentiality disfavours *doen*.
>
> Prediction 4 is not confirmed in our dataset. Lexical fixation does seem to affect the preferences for *doen* and *laten* to some extent. However, it must be added that the way in which this happens is complicated and a thorough understanding of these effects, especially the complicated interaction between 'lexical fixation' (`sig.lex.col`) and 'typical association with causative constructions' (`sig.sem.col`), requires further investigation.
>
> Prediction 5 seems to be mostly confirmed in our dataset, but needs further inspection. In the model with the interactions, infinitives which are typically

associated with causative constructions indeed seem to favour *doen*, but there are two interactions which complicate matters. In the context of transitive constructions as well as in the context of lexically fixed infinitives the general pattern (infinitives typically associated with causative constructions favour *doen*) is reversed and in the first of these contexts (transitive construction) the joint effect of typical association with causative constructions and transitivity even disfavours *doen* to a surprisingly large extent.

Prediction 6 is confirmed in our dataset, even though some of the more detailed expectations were incorrect. The distribution of *doen* and *laten* is indeed sensitive to lectal differences. And indeed Belgian origin of the data favours *doen*, as was predicted. However, we also formulated a specific informal register related sub-hypothesis in the beginning of this paper: within the Belgian sources, we expect more *doen* in informal registers than in formal ones. This was not confirmed: what we found instead, is that spontaneous speech (as opposed to prepared speech) disfavours *doen*, and it does so in Belgium as well as in The Netherlands.

Most of the predictions we derived from the (in)direct causation hypothesis (at least in the sense in which we interpreted it) were not confirmed by the case study. Therefore the assumptions on which the predictions were based should be reconsidered. The case study shows that the (in)direct causation hypothesis, when interpreted along the lines that were described in the introduction to this paper, is not tenable. This need not imply that the hypothesis should be abandoned entirely, but it does narrow down the number of legitimate interpretations of the hypothesis. But is there an alternative? Suggesting an alternative interpretation for the data basically means finding a framework that makes optimal sense of the various observations that follow from the statistical analysis. It would seem, then, that there are two features that characterize doen in comparison to laten.

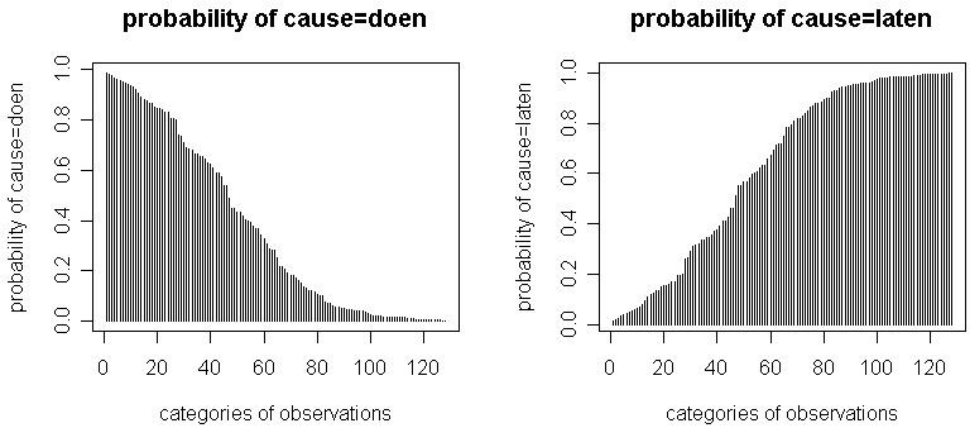
In the first place, *doen* exhibits a type of behavior that is typical of obsolescent or archaic forms. Its overall frequency is significantly lower than that of *laten*, which suggests that *laten* is the default form. Further, *doen* occurs more readily in formal and written registers (a type of language use that is likely to maintain expressions that have disappeared from everyday parlance), and it is more frequent in Belgian Dutch (which is generally more archaic than Netherlandic Dutch). In addition, it is not uncommon for obsolescent forms to continue their existence as lexical relics in idiomatic phrases: the importance of lexical fixation for the occurrence of *doen* seems to point in precisely that direction.

In the second place, *doen* is semantically specific, to the extent that it is preferred in cases of direct material causation: direct causation as indicated by the transitivity factor (when there is no causee, the preference for *doen* increases), and material causation as indicated by the animacy factor (the preference for *doen* increases with inanimate matrix subjects). Furthermore, if we assume that this direct material causation as referred to by *doen* is the prototypical core case of causation, it makes sense that infinitives that are typically 'causative-prone' (i.e.

that collocate semantically with causativity) increase the presence of *doen*.

A semantic characterization of *doen* in terms of direct material causation obviously comes close to the specific interpretation Kemmer and Verhagen originally gave of the (in)direct causation hypothesis, as discussed in section 2. Our data are consonant with their observations on the distribution of *doen*, but we do not think that this suffices to corroborate the (in)direct causation hypothesis as such - or in other words, we cannot say that the (in)direct causation hypothesis as a whole is an adequate framework for describing the distribution of *doen* and *laten.* There are two compelling reasons for this. First, the distribution of *doen* is not just determined by the semantic factor 'direct material causation', as we just saw. Second, while a particular type of direct causation does seem to play a role in the case of *doen*, it would be too simple to conclude that *laten* is therefore determined by indirect causation. *Laten*, in fact, appears to have a wider distribution than *doen,* in the sense that there is a wider range of contexts with a high probability for *laten* than there is for *doen*. Consider Figure 3, which charts the probabilities for encountering *doen* (left hand plot) or *laten* (right hand plot) for the 128 conditions that are defined by the possible combinations of predictor values. In each plot, the 128 conditions are ordered by decreasing probability for *doen*. For each individual context, the two plots are obviously complementary, but the size of the shaded area in the right hand panel is much bigger than that in the left hand panel: *laten* covers more contexts than *doen*.

Figure 3: Probabilities for *doen* or *laten* over categories of observations



The two characteristic features of *doen* may be brought together under the statement that *doen* is a marked form in comparison with *laten*: both in a formal, stylistic sense and in a semantic sense, *doen* appears to have a more restricted and more specific range of application than *laten.* But do these two forms of specialization point in the same direction ? The intriguing interaction between the lexical fixation effect and the semantic collocation effect suggests that they do not. The lexical fixation effect is part of the formal specialization of *doen*: obsolescent forms often survive in idioms and lexical freezes. The semantic

collocation effect, on the other hand, is representative of the semantic specialization of *doen*: if *doen* is typical for core cases of material causation, then the more 'causative-prone' a verb may be, the more it favours *doen.* The combination of lexical fixation and semantic collocation, however, does not boost the presence of *doen* (as would be the case if the two forms of specialization pointed in the same direction), but on the contrary drastically reduces the choice for *doen.* This suggests that the two factors that we just identified do not point in the same direction. They are complementary rather than conflicting, however: the interaction precisely suggests that in those cases in which you cannot predict *doen* on semantic grounds, it can be predicted on formal grounds, i.e. *doen* is predicted either by the fact that it preferentially patterns with 'causative-prone' verbs, or by idiomatization, but not by the two together. (If causative-prone verbs are lexicalized, they take *laten*, which is plausible: the presence of *doen* would not need to be lexicalized, since it can be predicted on semantic grounds.)

Needless to say, this alternative hypothesis is precisely that: a hypothesis that needs further scrutiny. We mentioned in the beginning of the paper that our preliminary investigation is not based on an exhaustive coding of the observations. For instance, we should try to operationalze the semantic factors that would be relevant for the Kemmer-Stukker-Verhagen interpretation of the (in)direct causation hypothesis, as suggested in section 2. The alternative interpretation formulated here calls for further data as well. If, for instance, *doen* is indeed an obsolescent form, we should be able to trace that development on the basis of diachronic materials. In any event, however, it will have become apparent that the distribution of *doen* and *laten* is governed by a more complex set of factors than the simple opposition of direct and indirect causation.

6. Conclusions and wider perspectives

Even though the results presented in this paper are not definitive, they are however clear enough to cast doubt on the (in)direct causation hypothesis - at least in the face value interpretation of the hypothesis that we started from and which, to repeat, does not coincide with the interpretation originally suggested by Kemmer and Verhagen. Starting from a set of 3975 cases of *doen* or *laten* extracted from the Spoken Dutch Corpus, we performed a stepwise logistic regression analysis incorporating a series of factors which on the basis of the (in)direct causation hypothesis were predicted to affect the choice between the use of either *doen* or *laten* in specific ways. The results show that most of these predictions are falsified, and that it will therefore be necessary to pursue a different basic hypothesis about the causes for choosing either *doen* or *laten*: we have suggested that as a causative verb, *doen* is an obsolescent form with a tendency towards semantic and lexical specialization. But consonant with the idea of an empirical cycle that is part and parcel of the scientific method, this reinterpretation is merely a hypothesis for further testing.

We introduced our causatives case study as an example of how the scientific method can be used in linguistics. The 'scientific method', needless to say, is the ap-

proach to scientific investigation in which the empirical testing of hypotheses is paramount: systematic data gathering on the basis of observation or experimentation yields material that may be used to falsify predictions derived from a theoretical hypothesis. Because this approach to scientific enquiry is not as dominant in linguistics as it is in other behavioral disciplines, like sociology and psychology, we may now conclude by summarizing a few central aspects of empirical research as meant by the scientific method, and as illustrated by our case study. (This passage is an elaboration of a number of remarks made in Geeraerts 2006.) What are the main features of empirical research?

1. Empirical research is *data-driven*. You cannot easily draw conclusions from single cases and isolated observations, and the more data you can collect to study a particular phenomenon, the better your conclusions will get.

2. Empirical research in linguistics may be *observational or experimental;* there is a complementarity between both approaches. The research data may come from different sources: they may be collected as they exist (as is the case in corpus research), but they may also be elicited by doing experimental research, or by doing survey research. As applied to language, the mutual advantages of observational versus experimental research are clear: observational research (viz. corpus research) allows you to study language in a natural and spontaneous state; but experimental research, by contrast, may give you a better control over specific variables, as when they are underrepresented in the corpus.

3. Empirical research involves *quantitative methods*. In order to get a good grip on the broad observational basis of elicited and/or non-elicited data, investigators need techniques to come to terms with the amount of material involved. Specifically, they will need statistical tests to determine whether specific observations might be due to chance or not.

4. Empirical research crucially hinges on asking the right questions, or in other words, on the *formulation of hypotheses*. No perception could be more misguided than to think that once you have your database of elicited or non-elicited observations, the conclusions will arise automatically and purely inductively from the data. On the contrary, the only conclusions you will be able to draw are the ones that relate to hypotheses you have formulated and tested – so that will be the investigator's first task. Another way of saying this is that empirical research necessarily combines inductive and deductive reasoning: on the one hand, you work in a bottom-up way from data to hypotheses, but on the other hand, those hypotheses will also be derived top-down from the theoretical perspective you adopt in thinking about your data.

5. Empirical research requires the *operationalization of hypotheses*. It is not sufficient to think up a plausible and intriguing hypothesis: you also have to formulate it in such a way that it can be put to the test. That is what is meant by 'operationalization': turning a hypothesis into concrete predictions that can be tested against the data. In most empirical research in linguistics, it is questions of operationalization that require all the ingenuity of the researcher – and most of his or her time, because getting the relevant data and measurements is not an automatic process.

6. Empirical research involves an *empirical cycle* in which several rounds of data gathering, testing of hypotheses, and interpretation of the results follow each other. Just like it is misguided to think that empirical, data-driven research automatically gives one all the answers, it is misguided to think that it immediately gives one the final answer. The empirical cycle as such, in fact, does not constitute a straightforward march towards the truth, because negative results may be interpreted in different ways. If a prediction is not borne out, at least two kinds of interpretation suggest themselves: the original hypothesis (or the broader framework in which it is couched) may be wrong, but in principle, it could also be the case that our operationalization of the hypothesis was not adequate. The assumption may be wrong, or our way of testing the assumption may be inappropriate - but the consequences in either case are largely different. Empirical research seeks maximal objectivity, but it is in no way a mechanical procedure that inevitably leads to a single possible result. That is not the way it happens in the hard sciences, and it is not the way it happens in the study of language either.

7. Empirical research does not rule out *creativity and intuition*. To the undiscerning eye, the ideal of scientific objectivity would seem to banish the investigator as a subject from the investigation, but a closer look makes clear that ingenuity and interpretative insight are indispensable features of the empirical cycle. Hypotheses translate an intuitive understanding into operational predictions; finding the right operationalization rests on inventiveness as much as on expertise; and processing the results of the empirical cycle requires creative imagination. Empirical research does not lower the demands on the subjective skills of the researchers; it only raises the criteria for the objective validity of their claims.

References

d'Andrade, Roy. 1987. A folk model of the mind. In Dorothy Holland & Naomi Quinn (eds.), *Cultural Models in Language and Thought* 112-148. Cambridge: Cambridge University Press.

Den Boon, Ton & Dirk Geeraerts. *Van Dale Groot woordenboek van de Nederlandse taal*, 14th edition. Utrecht/Antwerpen: Van Dale Lexicografie.

De Clerck, Walter. 1981. *Nijhoffs Zuidnederlands Woordenboek*. 's Gravenhage/Antwerpen: Martinus Nijhoff.

De Sutter, Gert, Dirk Speelman & Dirk Geeraerts. 2005. Regionale en stilistische effecten op de woordvolgorde in werkwoordelijke eindgroepen. *Nederlandse taalkunde* 10: 97-128.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1): 61-74.

Geeraerts, Dirk. 2006. Methodology in Cognitive Linguistics. In Gitte Kristiansen, Michel Achard, René Dirven & Francisco Ruiz de Mendoza Ibañez (eds.), *Cognitive Linguistics: Current Applications and Future Perspectives* 21-49. Berlin/New York: Mouton De Gruyter.

Geeraerts, Dirk, Stefan Grondelaers & Dirk Speelman. 1999. *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen.* Amsterdam: Meertens Instituut.

Gibbs, Raymond W. 2007. Why cognitive linguists should care more about empirical methods. In Monica Gonzalez-Marquez, Irene Mittelberg, Seana Couslon & Michael J. Spivey (eds.), *Methods in Cognitive Linguistics* 2-18. Amsterdam/Philadelphia: John Benjamins.

Glynn, Dylan. In press. Polysemy, syntax, and variation. A usage-based method for Cognitive Semantics". In Vyvian Evans & Stephanie Pourcel (eds.), New Directions in Cognitive Linguistics. Amsterdam: John Benjamins.

Gries, Stefan Th. 2006. Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54:191-202.

Heylen, Kris. 2005. A quantitative corpus study of German word order variation. In Stephan Kepser & Marga Reis (eds.), *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives* 241-264. Berlin, Mouton de Gruyter.

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.

Rietveld, Toni and Roeland van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behavior*. Mouton De Gruyter: Berlin.

Schuurman, Ineke, Machteld Schouppe, 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*. Cambdrige: MIT Press.

Talmy, Leonard. 2007. Introspection as a methodology in linguistics. Plenary lecture presented at the 10th International Cognitive Linguistics Conference, Krakow, July 2007.

Tuggy, David. Schematicity. In Dirk Geeraerts & Hubert Cuyckens (eds.), *The Oxford Handbook of Cognitive Linguistics* 82-116. New York: Oxford University Press.

Tummers, José, Dirk Speelman & Dirk Geeraerts. 2005. Inflectional variation in Belgian and Netherlandic Dutch: A usage-based account of the adjectival inflection. In Nicole Delbecque, Johan van der Auwera & Dirk Geeraerts (eds.), *Perspectives on Variation. Sociolinguistic, Historical, Comparative* 93-110. Berlin/New York: Mouton de Gruyter.

Tummers, José, Kris Heylen and Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1: 225-261.

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.

Verhagen, Arie & Suzanne Kemmer. 1997. Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics* 27, 61-82.