FACULTEIT LETTEREN

SUBFACULTEIT TAALKUNDE

KATHOLIEKE
UNIVERSITEIT
LEUVEN

Doe wat je niet laten kan: A usage-based analysis of Dutch causative constructions

Proefschrift ingediend tot
het behalen van de graad van
Doctor in de Taalkunde

Natalia Levshina

Promotor:
Prof. dr. Dirk Geeraerts
Co-promotor:
Prof. dr. Dirk Speelman

November 2011

Acknowledgements

This project was funded by a grant from the Flemish Research Foundation – FWO (G.0330.08). I'm grateful to my supervisors Professors Dirk Geeraerts and Dirk Speelman for giving me a chance to prove myself and selecting me for this job. I also thank the Department of Linguistics and the Faculty of Arts of the Katholieke Universiteit Leuven for excellent working conditions. Here at the KULeuven I've had a privilege to be be inspired by and learn from outstanding professionals in a unique cultural and academic environment. I should also mention with gratitude my previous teachers at Pskov Arts Lyceum, Pskov Pedagogical University and the Department of General Linguistics of Saint Petersburg State University, who encouraged my creativity and interest in linguistics. My warmest thanks go to the (ex-)colleagues in the Quantitative Lexicology and Variational Linguistics research unit for practical and emotional help whenever it was needed and for a great company. I'm especially grateful to Kris Heylen for inspiration and support at all stages of the project, Tom Ruette for the opportunity to use his brand-new Usenet corpus, and Alena Anishchanka, Jack Grieve and Eline Zenner for their constructive feedback when this dissertation was shaping up. I want to thank all native-speaking colleagues for helping me to decipher some examples of non-elicited use of Dutch. All remaining errors are, of course, solely mine. I'm also deeply touched by the unconditional support and never-ending encouragement from my family and close friends – during the last four years and always.

Contents

Introduction..1

Chapter 1. What is (constructional) meaning and how to describe it....5

1.1. Meaning as a cognitive phenomenon.......................................6

1.2. Meaning as a social phenomenon...14

1.3. Empirical corpus-based methods of modelling semantics......19

1.3.1. An overview of existing methods in Cognitive Semantics...19

1.3.2. Corpus-based distributional methods.......................24

1.4. Summary..29

Chapter 2. Variation of the Dutch causative constructions with *doen* and *laten*..30

2.1. Form and function of Dutch causative constructions with *doen* and *laten*..30

2.2. The constructional network of *doen* and *laten*........................37

2.2.1. Transitive and intransitive constructions 38

2.2.2. Marking of the Causee 39

2.2.3. Causeeless constructions 40

2.2.4. Lexically specific constructions with *doen* and *laten* 43

2.3. Dutch causative constructions: diachronic and synchronic variation..46

2.4. Summary..50

Chapter 3. Data and method..52

3.1. A theoretical and methodological rationale of the approach...52

3.2. Corpus design..56

3.3. Variables..59

3.3.1. Local contextual variables.......................................60

3.3.2. Global contextual variables......................................61

3.4. Quantitative analysis procedure...61

3.5. Summary..69

Chapter 4. Semantic and lectal variation of the causative construction with *doen*..71

4.1. Intracategorial salience of exemplars of *doen* 71

4.2. The exemplar space of *doen*...73

4.3. Hierarchical cluster analysis of the exemplars of *doen*...........80

4.4. Lectal variation of *doen*..92
4.5. Summary...99

Chapter 5. Semantic and lectal variation of the causative construction with *laten*...101
5.1. Intracategorial salience of exemplars of *laten*.......................101
5.2. The exemplar space of *laten*..102
5.3. Hierarchical cluster analysis of the exemplars of *laten* 108
5.4. Lectal variation of *laten*...117
5.5. Summary...124

Chapter 6. *doen* vs. *laten*: division of semantic and lectal labour......126
6.1. Cue validity of exemplars of *doen* and *laten*.........................126
6.2. The common exemplar space of *doen* and *laten*...................128
6.3. Distinctive features of *doen* and *laten*..................................137
6.3.1. Exploratory analysis...137
6.3.2. Confirmatory multivariable analysis......................139
6.4. Lectal differences in the division of semantic labour between *doen* and *laten* 150
6.5. Summary...155

Chapter 7. Discussion and perspectives..157
7.1. All sorts of salience...157
7.2. Is it done with *doen*?..161
7.3. Towards an operationalizable semantic theory......................164

Appendix 1. Local contextual variables...167

Appendix 2. Random forests for six lects..175

References..178

Introduction

Language is a semiotic system. When using it for communication, we rely on more or less conventional links between form and meaning, which are expected to be shared by most speakers in a linguistic community. Many linguists (e.g. Langacker 1987; Goldberg 1995; Croft 2001) believe that form-meaning pairings are observed at most levels of language structure, including syntax. In this context, it is surprising that linguists have not reached an agreement about how to study meaning, let alone developed a common set of tools and practices for semantic description. The empirical (corpus-based or experimental) approaches to semantics are not yet mainstream. Many semanticists still prefer the subjective introspective method as the main gateway to meaning (e.g. Talmy 2007).

The easy accessibility of meaning via introspection, however, is an illusion (see Geeraerts 2010b). Grammatical constructions provide a good example. For instance, what is the meaning of the Dutch causative constructions with *doen* "do"and *laten* "let", which are the objects of the present study? What are their common semantic functions, and which are unique? In order to grasp the semantics of the constructions, one will invariably start thinking of examples and specific constructional patterns. But how can one be sure that all important patterns are taken into account? And what is the weight of these patterns in the overall semantic structure? Is there regional or situational variation in the use of the constructions? To answer these questions, we need objective evidence.

This evidence can be collected in abundance from existing large corpora. Unfortunately, current corpus methods frequently lack interpretability in semantic terms. Ardent adepts of corpus methods sometimes present their arcane statistical models with little if any theoretical generalization, leaving the unconverted with a "so what?"

feeling. A present-day semanticist is too often trapped between Scylla of unverifiable intuitions and Charybdis of uninterpretable corpus counts.

The novel approach proposed in this study aims at maximizing both objectivity and interpretability. It is based on linguist-friendly intuitive visual representations of linguistic categories as constellations of their exemplars (understood as unique instances) in a semantic space. The 'senses' are defined as clusters of similar exemplars. The semantic and visual distance between the exemplars is established on the basis of their semantic and other features. A range of standard multivariable techniques is used to obtain linguistically relevant information, such as the most salient dimensions of semantic variation, semantic scope of linguistic units and their overlap, clusters of usage patterns organized in a hierarchical network, or autonomy and entrenchment of the 'senses'. The approach represents thus one of the first fully bottom-up models of semantic structure. Additional techniques (logistic regression with mixed effects, conditional inference trees and random forests) are used to establish whether the results can be extrapolated to the entire population of the constructions under study.

I hope that this approach will help to bridge the gap between the theory of Cognitive Linguistics and the corpus-based approaches to meaning (cf. Arppe et al. 2010). Unfortunately, most of the theoretical notions and hypotheses in Cognitive Linguistics are difficult to operationalize in a testable way. This means that they are metaphysical in Popper's sense (Stefanowitsch 2010). The ultimate – and the most ambitious – goal of the present study is to interpret the results of the quantitative analyses according to the principles of the usage-based approach to language and suggest some new testable hypotheses, thus contributing to a new empirically oriented semantic theory (cf. Stefanowitsch 2010).

Another novel aspect of this study is the integration of the semasiological (from form to function) and onomasiological (from function to alternative forms) perspectives, which is not common in Cognitive Linguistics in general (Geeraerts et al. 1994 and Glynn 2007 are the few known exceptions) and especially in studies of grammatical constructions. Most studies either explore an 'alternation' of near-synonymous constructions, or focus on one construction. The present work is meant to reveal which aspects of meaning become more prominent with the shift of perspective and which remain the same.

The approach proposed here also allows for an intuitive interpretation of lectal variation in the semantic structure of linguistic categories. This study attempts to capture the variation in two national varieties of Dutch and three different registers of communication. Since some varieties are believed to have retained more archaic features than the others, this will enable me to test several existing hypotheses about the development of the constructions using the synchronic evidence as a 'time machine'. In this way, the study is a contribution to variational linguistics, which is still biased towards the units that convey as little conceptual content as possible – to a large extent, it seems, due to the lack of tools that allow a sociolinguist to control for meaning (cf. Geeraerts 2010a). With this approach, I hope to extend the inventory of tools that can be used to disentangle intricately connected semantic and lectal factors that influence language variation and change.

The thesis has the following structure. First, I introduce the theoretical and methodological background of the study. Chapter 2 discusses variation in structure and semantics of the causative constructions with *doen* and *laten*, as described in previous research. I also summarize the known facts about the geographic and situational variation in the use of the constructions. Chapter 3 introduces the corpus data used in this study and describes the procedure of the analysis as a sequence of steps. Next, I approach the constructions from the semasiological perspective and model the semantic dimensions and polysemy of *doen* (Chapter 4) and *laten* (Chapter 5) in Dutch in general and in the above-mentioned varieties. I also use historical evidence and child language acquisition data to support the interpretation. Chapter 6 adds the onomasiological perspective to the study. It discusses the distinctive semantic features of *doen* and *laten* and focuses on the geographic and situational factors that influence the division of labour between the two constructions. In Chapter 7, I discuss the results of the quantitative analyses from the general conceptual usage-based perspective. I interpret the findings with regard to the existing hypotheses about language variation and change, and suggest new ones.

Chapter 1. What is (constructional) meaning and how to describe it

Before describing the semantics of the Dutch causative constructions, it is necessary to understand what meaning is and how it can be studied. These questions are answered here in line with empirical Cognitive Semantics – a dynamic subfield of Cognitive Linguistics,[1] which describes, explains and predicts semantic phenomena in natural language with the help of empirical (corpus-driven and experimental) methods. Cognitive Linguistics is a usage-based framework (Langacker 1987; Barlow and Kemmer 2000), which assumes that knowledge of language is both reflected in and shaped by language usage. This approach discards the structuralist and generativist oppositions "*langue – parole*" and "competence – performance" as no longer relevant. From the external perspective of demarcation of linguistics as a discipline, there is no principled difference between linguistic and extralinguistic knowledge. Use of language involves a complex interaction of various conceptual, social, cultural, processing-related and other factors, which should all be taken into account in a linguistic model. As a consequence, language research is necessarily multifactorial and interdisciplinary. From the internal perspective, the usage-based approach leads to the disappearance of traditional distinctions within linguistics, such as the distinction between lexis and grammar. Linguistic units at most levels – from morphemes to abstract syntactic constructions – are learned and processed as form-function pairings, which differ only in compositionality and schematicity

[1] The terms Cognitive Linguistics and Cognitive Semantics are written with capital letters to refer to a specific framework in cognition-oriented studies of language, which originated in the 1980s (e.g. Lakoff 1987, Langacker 1987) as a reaction to the generativist paradigm.

(e.g. Langacker 1987; Goldberg 2006). Situated on the borderline between lexicon and grammar, the Dutch causative constructions with the auxiliaries *doen* and *laten* are a perfect object for such an integrative approach.

The first two sections of this chapter outline the most important aspects of meaning, which determine the design of the present study. The third section describes existing methods in semantic research, focusing on the family of empirical corpus-based distributional methods, which the present study belongs to.

1.1. Meaning as a cognitive phenomenon

Cognitive Linguistics treats linguistic meaning as a psychological, or cognitive phenomenon – a concept or conceptual structure associated with the given linguistic form (e.g. Langacker 1987). Speakers verbalize their experience by choosing the linguistic forms associated with specific conceptual structures in the mind. A very simplified representation of the processes of verbalization and comprehension is shown in Figure 1.1. This scheme displays relatively stable links between experience, concepts and linguistic forms in the mind of an individual speaker. The figure can be also interpreted as a generalized representation of linguistic and conceptual (encyclopaedic) knowledge shared by members of a linguistic community. Because the correspondences between experiences, conceptualizations and linguistic forms are never one-to-one, less typical links, which are less frequently activated, are shown here, as well. They are represented as dashed arrows. The secondary Concept – Form links are displayed because not all variation in language is due to purely conceptual reasons. For instance, one form may be chosen over another due to information-processing factors (e.g. if the alternative is too long), or sociolinguistic reasons (e.g. if the other variant is socially stigmatized).[1]

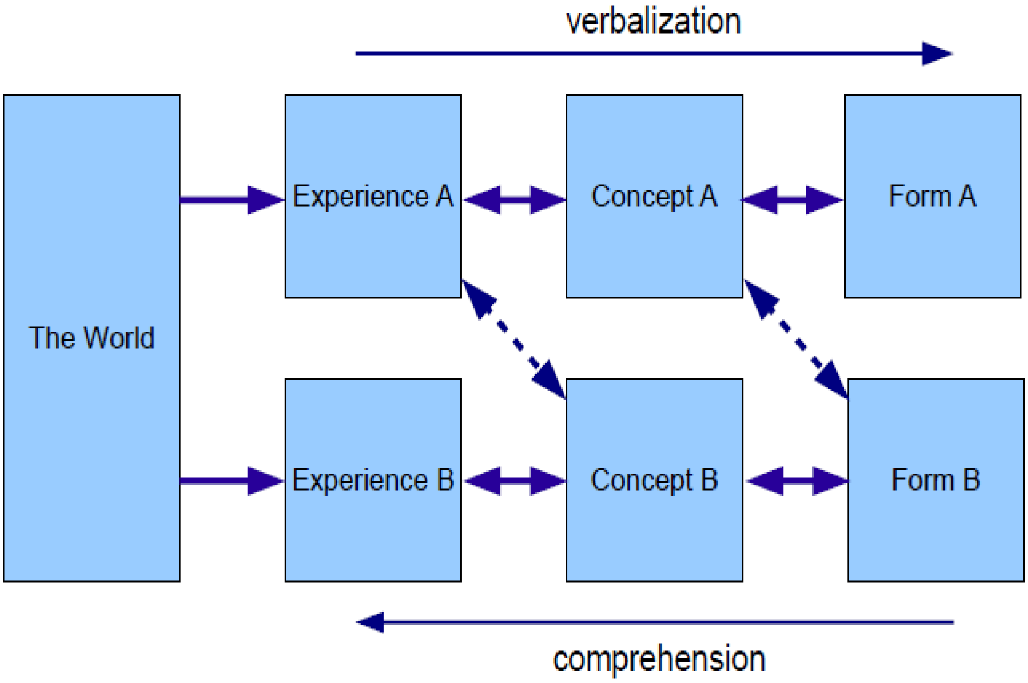


Figure 1.1. Cognitive processes of verbalization and comprehension.

If we accept the 'cognitive commitment' of Cognitive Linguistics (Lakoff 1990), we should try to take into account the cognitive processes that are described in psychological studies of concepts and categories. According to most of them, concepts are formed by the traces of specific objects and situations associated with a given category (the category's extension, according to the traditional semantic terminology) and their properties (the intension) in the memory. The links between experience and concepts are of primary importance in linguistic and non-linguistic categorization. Psychologists have been exploring what makes some objects and situations serve as better representatives of a category than other objects and

[1] Strictly speaking, the sociolinguistic information of this kind is conceptual, too, but whether this information is stored together with the 'semantic' information, or separately, is an open question.

situations. For instance, it was shown that subjects' ratings of goodness-of-membership of a category member correlate with its family resemblance, operationalized as the number of shared features with the other members of the category (Rosch and Mervis 1975). For example, an orange is a better representative of the superordinate category FRUIT than an olive because it shares more features with the other members of the same category.

Researchers are also interested in the question what determines the chances of a concept to categorize one and the same chunk of experience. For instance, the category FRUIT may be more salient for designating an orange than the category VEGETABLE because the object shares more properties with other fruits than with vegetables. Another reason can be a higher degree of basicness of the winning category (Rosch et al. 1976; Berlin 1978). For instance, in most everyday contexts a butterfly will be called a butterfly, and not with its more specific name, e.g. *Erebia Embla*, unless this is a forum of entomologists or a novel by Nabokov. The additional information that the highly specific category contains is of little use in everyday life. On the other hand, the superordinate category INSECT will not be chosen, either, because of its low informativity.

One of the main questions in psychological studies of categories and concepts is the degree of abstraction involved in categorization choices and mental representations of categories (Vanpaemel and Storms 2008). According to the Prototype Theory (Rosch 1975; Rosch and Mervis 1975; Rosch 1978), people form highly abstract representations (prototypes), generalized over all instances of a category, and then compare every new candidate with the prototype. In other words, the Prototype Theory highlights the importance of the semantic intension. On the other extreme is the Exemplar Theory (Medin and Schaffer 1978; Nosofsky 1986), which focuses on the extension. It claims that people store only traces of specific exemplars in the memory, without any abstract representations. High similarity of a stimulus to a few exemplars of a category is more important for categorization than moderate similarity to many exemplars of the same category, contrary to what one might expect according to the Prototype Theory. More recently, new models of categorization have been developed (e.g. Varying Abstraction Model in Vanpaemel and Storms 2008), which demonstrate that people may use some intermediate forms of abstraction to represent natural language categories. Still, little is known about how these abstractions are formed, in particular in natural language categories.

Historically, the Prototype Theory has been the predominant view in Cognitive Semantics (see an overview in Geeraerts 2010c: 183–203). Attempts have been made to extend the prototype approach from the natural kinds and artifacts, which are traditionally used in psychological experiments, to more abstract semantics of function words, verbs and syntactic categories (e.g. Brugman 1983; Lakoff 1987; Taylor 1989; Pulman 1993: Ch. 5; Geeraerts 1998). However, most linguistic semantic studies of this kind lack a solid empirical ground. As such, they are a mere "exercise in speculative psychology" (Stefanowitsch 2010: 374), and the scientific status of their results is problematic. More recently, attempts have been made to model prototypicality effects in lexical and constructional semantics with the help of quantitative techniques applied to large-scale corpus data (Gries 2003; Gries 2006), although the relationships between prototypicality and various corpus frequency measures are not yet entirely clear (Gilquin 2006; Schmid 2010).

The Exemplar Theory has had a smaller impact on Cognitive Linguistics so far, although this theory has been quite popular in usage-based approaches to language variation and change (Pierrehumbert 2001; Bybee 2006). Yet, there are a few linguistic studies that apply the principles of the Exemplar Theory to constructions (Bybee and Eddington

2006; Zeschel 2010). For instance, Bybee and Eddington's (2006) research of several Spanish constructions *become* + Adjective demonstrates that the similarity of an adjective to a few highly frequent adjectives in the Adjective slot increases the acceptance rate of the contexts with this adjective.

It is necessary to mention, however, that linguistic exemplars in these studies are treated as low-level schemata – partly lexicalized instantiations of a construction (e.g. a combination of a specific verb of becoming and a specific adjective). Bybee herself writes that exemplars “are built up from tokens of language experience that are believed to be identical” (Bybee 2010: 7). But it is not clear how and under what conditions this abstraction process takes place. For instance, why are the exemplars of the above-mentioned *become*-construction defined as pairs of a verb and an adjective, and not as trios of a 'becomer', a verb, and an adjective, or as pairs of a specific verb form and an adjective, or in any other way? Note that psychologists define exemplars very differently. Some treat them as unique instances, and some as subordinate abstract categories, e.g. the subcategories SPARROW and PENGUIN can be called exemplars of the category BIRD (see an overview in Storms, De Boek and Ruts 2000 and Murphy 2002: 58–60). In the present study, I understand exemplars as individual tokens. Yet, low-level schemata (exemplars in Bybee's sense), as well as more schematic constructions emerge in the form of clusters of similar instances.

Unfortunately, not much empirical evidence is available about the level of abstraction in mental representation of constructions, although Goldberg (2006) with the help of experimental and corpus evidence shows that generalizations (schematic constructions, e.g. the transitive construction) do play an important role in acquisition and use of constructions. The present study therefore assumes that both generalizations and exemplar-specific information are relevant for a semantic description.

Another fundamental aspect of language semantics is the difference between the inter- and intracategorial perspectives (in semantic terms, the onomasiological and semasiological views). The most typical linguistic phenomena associated with these perspectives are polysemy and synonymy, respectively. The senses of a polysemous unit correspond to overlapping conceptual structures in the conceptual 'store' of a speaker (Murphy 2002: 391). Conversely, synonymy means that a part of conceptual structure is shared by several linguistic categories. The rarity of full synonymy in language is explained by the phenomenon of preemption: “one cannot give a word a meaning that another word already has” (Murphy 2002: 409), which ensures that the conceptual system functions economically. The phenomenon of preemption implies that the conceptual structure of one linguistic unit is affected by the existing alternatives (cf. Goldberg 2002: 349). There is evidence that the semantic structure of a linguistic unit is shaped by the behaviour of its 'competitors', especially in historical semantics (e.g. Blank 1999; Rastier 1999). This is why both perspectives should be taken into account, especially when studying variation and change.

However, attempts to model meaning from the two perspectives, taking into account both the internal structure of a linguistic category and the division of labour between similar categories, are very rare and belong to the domain of lexicology (Geeraerts et al. 1994, and, more recently, Glynn 2007). It is an interesting fact that during its academic history, linguistic semantics has been fluctuating like a pendulum between intracategorial and intercategorial relationships (see Geeraerts 2010c, who charts its trajectory in detail). Whereas structuralists concentrated predominantly on synonymy and other relations between linguistic units, the best-known descriptive Cognitive Linguistics studies focused mainly on polysemy (e.g. Brugman 1983; Lakoff 1987). The contemporary usage-based studies of constructions, however, (e.g. Grondelaers et al. 2002; Gries 2003; Gries and Stefanowitsch 2004; Heylen 2005; De Sutter 2009), display interest in 'alternations', or constructional near-synonymy. Because this interest is in a way a relic of the generative tradition, Goldberg argues that constructions should be studied in their own right (Goldberg 2002; see also Colleman 2010), although she does not dismiss the onomasiological intercategorial approach, either, due to the reasons described above. It seems that disregarding one or the other perspective can have a negative effect on the interpretation of the results in terms of 'prototypes' or 'salience' because, as some psychological studies show, the inter- and intracategorial salience effects (e.g. family resemblance and cue validity) do not always correlate (e.g. Ceulemans and Storms 2010).

An important note concerns organization of semantic structure. Some Cognitive Semanticists represent polysemous categories as radial networks of extensions from the semantic core – the 'prototype' (e.g. Brugman 1983; Lakoff 1987). The main units of analysis are gestalt-like conceptual entities, which are traditionally represented as discrete nodes in the network, although it is frequently assumed that these 'focal senses' exist in a semantic continuum (Brugman 1983). Some others (e.g. Geeraerts 1998) focus more on the dimensions which form the extensions than on the specific senses. In the present study I combine both perspectives, operationalizing senses as more or less distinct clusters of similar exemplars in a semantic space. These clusters differ along several conceptual or other dimensions and emerge as a result of quantitative analysis.

Finally, it is necessary to make some qualifications about constructional semantics. In the present study, constructions are understood as a unity of form and conceptual contents, with additional information about the lectal and pragmatic usage patterns. As mentioned above, constructional semantics is not fundamentally different from lexical semantics, which has a longer tradition in linguistics. Although constructional semantics tends to differ from lexical semantics by being more abstract and compositional, one can expect to find the same phenomena: polysemy and synonymy, prototypicality and exemplar effects. Goldberg (1995), for instance, shows that most argument structure constructions are polysemous, with a central sense and extensions based on metaphorical or other links. For example, the central sense of the English ditransitive construction is the successful transfer between a volitional agent and a willing recipient (*She gave him a book*). One of the extensions is the metaphorical transfer of information from the stimulus to the recipient (*She told him the news*) (Goldberg 1995: Ch. 6). An important peculiarity is that the semantic relationships between the 'senses' of one construction are normally motivated by and reflected in their constituents (e.g. Croft and Cruse 2004: 274). These specific senses are therefore associated with a formal pattern, constituting a lower-level schema.

To summarize, a cognitively plausible semantic description of one or several linguistic categories (lexemes or constructions) should take into account the following aspects:

- prototypicality of specific exemplars (similarity to the abstract prototype, understood as a generalization over all exemplars), and exemplar effects (similarity to specific exemplars);
- semasiological (intracategorial) and onomasiological (intercategorial) perspectives on the meaning of the linguistic units, which correspond to family resemblance and cue validity of the exemplars;

- salience of individual exemplars or their clusters, and structural weight of specific conceptual features, or cues (cf. structural salience in Geeraerts 2006 [2000]).
- for constructions, it is also important to pay attention to the associations between different parts of conceptual structure and specific formal pattens, which together constitute low-level schemata.

In this study I propose an approach that can meet these requirements and demonstrate how it works, applying it to the Dutch causative constructions with *doen* and *laten*.

1.2. Meaning as a social phenomenon

According to the usage-based approach to language, the speaker's linguistic knowledge is fully determined by the linguistic input (s)he is exposed to, and the individual general cognitive abilities that are involved in processing this input. This is why one can expect substantial variation across individual speakers and linguistic communities, whose linguistic systems and subsystems are frequently referred to as *lects* – an umbrella term for idiolects, dialects, regiolects, sociolects, etc. Salient linguistic differences between communities become markers and stereotypes (Labov 1971), signalling group membership and taking part in the speaker's identity construction (Eckert 2008). In this case, the social dimension of linguistic variation becomes a part of the conceptual system.

This integration is easily interpreted within the conceptual framework, which assumes that the speakers store not only the concepts associated with the linguistic forms, but also knowledge about the situations in which these forms are used, including the social characteristics of the interactors, medium, goal and other interaction-related features – both situational and culturally entrenched ones. All this knowledge, which forms a part of the speaker's social cognitive abilities, enables him or her to use language as a coordinating device in joint actions (Clark 1999). While doing so, the speaker and the hearer take into account not only each other's actions and intentions, but also the models of each other's minds, including linguistic competence. In other words, they create cognitive and linguistic portraits of each other. Some of the knowledge necessary for creating these portraits comes from the actual situation; some can be predicted on the basis of the speakers' knowledge about the world, language varieties and communities (Croft 2007). Language change and variation can be seen then as the unintended result of this intentional individual behaviour, which can be metaphorically described as an “invisible hand” (Keller 1994).

Adding the variational dimension to the model of verbalization of comprehension suggested above, we can hypothesize that lectal variation in the use of a specific linguistic construct – for instance, differences in the frequency of a construction in two communities – can be caused by the following factors:

(i) differences in the experience of the speakers in the two communities, who have different chances to be exposed to specific situations or things. For instance, a simple Google search shows that the relative frequency of the word *kangaroo* is 8.5 higher in the sites with Australian domains than in those from New Zealand, whereas the word *kiwi* occurs 14 times more frequently in the New Zealand sites than in the Australian ones. A more complex example concerns the higher frequency of pronominal recipients in a corpus of telephone conversations than in a collection of news and financial reportage from *The Wall Street Journal* because of the obvious communicative differences between the registers. Bresnan et al. (2007) found that this difference explains the higher relative frequency of the double object dative in comparison with the *to*-dative in the spoken data because the former construction is more frequently used with pronominal indirect objects;

(ii) differences in categorization of similar referents due to linguistic or encyclopaedic variation in the categories and their organization. For instance, speakers of American English tend to categorize a piece of furniture for one person, with back and arm supports, as a chair, whereas British English speakers often prefer to refer to the same object as an armchair. This means that for the former the category CHAIR includes armchairs, whereas for the latter the categories are (more) distinct;

(iii) differences in the linguistic 'labels' attached to similar concepts or conceptual structures. This type of variation can be exemplified by sociolinguistic variables. For instance, a famous shibboleth that can help to tell a Moscovite from a person from Saint Petersburg is the word for a doughnut. Russian speakers from Moscow call it a *pončik*, whereas those from Saint Petersburg use the word *pyška*. If a Russian speaker from Saint Petersburg is aware of the shibboleth and asks for a *pončik* in a Moscow bakery, (s)he manipulates the linguistic labels for a pragmatic goal, i.e. buying the doughnut without getting into unnecessary linguistic negotiations.

These cases, which can be treated as the main potential sources of variation, are presented in Figure 1.2. Needless to say, disentangling these sources of variation is a very difficult task (cf. an attempt in Levshina et al., Submitted). Several sources of variation may interact with one another. For instance, a semantic change of a category towards abstractness, as in (ii) can make the category applicable for a broader range of situations and therefore increase its frequency, as in (i), which in turn can trigger further semantic change (cf. Landsbergen et al. 2010). One can imagine a case when the more frequent and abstract category also becomes more sociolinguistically and pragmatically neutral, which will affect the language at the level of (iii).

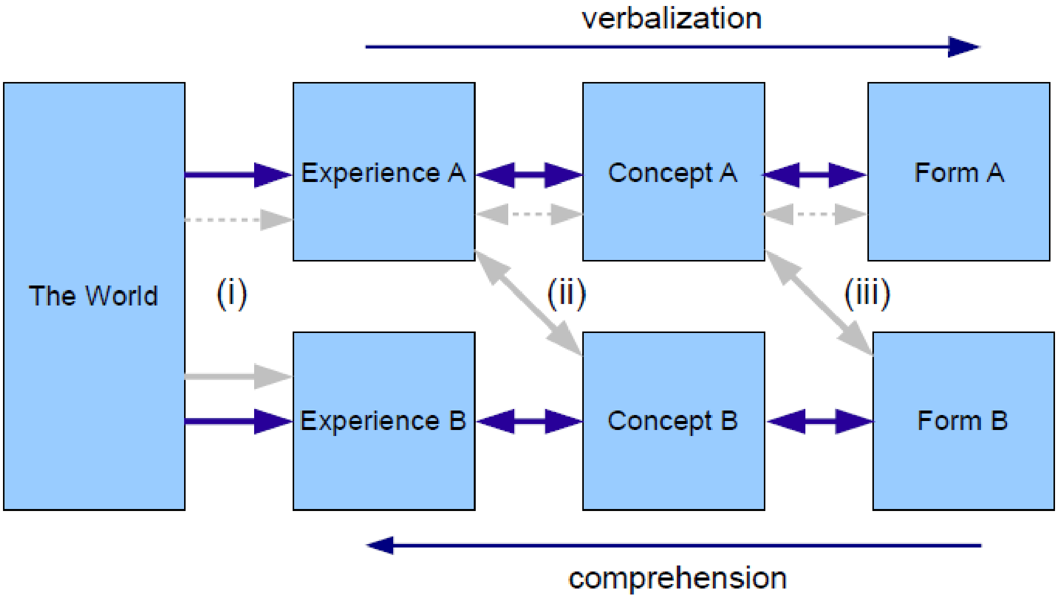


Figure 1.2. Main sources of linguistic variation. Darker links: 'default' Lect 1, lighter links: potential differences that can be found in Lect 2. (i)-(iii): sources of variation as listed above.

This model bridges the gap between the mental and the social, as well as between knowledge and behaviour. This is the aim of Cognitive Sociolinguistics – an interdisciplinary field that studies the interaction of semantic and social (understood broadly, including geographical, cultural and historical) variation in language perception and production (e.g. Kristiansen and Dirven 2008; Geeraerts et al. 2010). The studies show that this interaction is pervasive. It ranges from the historic and cultural grounding of specific metaphorical patterns, e.g. ANGER IS HEAT, in Geeraerts and Grondelaers (1995), to folk's mental models of linguistic varieties (Berthele 2010).

From the semantic perspective, the sociocognitive interaction can be exemplified by different constructions preferred for designating similar referents across language communities, or by different meanings associated with one linguistic form in two lects. Such crisp differences, though, should be less common than probabilistic variation, which can manifest itself in varying prototypicality of similar exemplars of the same category, or lectal differences in onomasiological salience of semantically similar categories (see Geeraerts et al. 1994). However, most sociocognitive studies focus on interlectal difference in the cue validity of specific semantic features that determine the speaker's choice from a set of alternatives (e.g. Grondelaers et al. 2002; Bresnan and Hey 2008; Speelman and Geeraerts 2009; Bresnan and Ford 2010, Szmrecsanyi 2010). For instance, Bresnan and Hey (2008) show that the effect of animacy on the chances of the ditransitive construction vs. the *to*-Dative is higher in New Zealand English than in the American variety.[1]

At the same time, there are critical voices, e.g. Gries and Divjak (2010), who question the effects that the choice of a lectally specific corpus can have on a corpus-based semantic description (see also Stefanowitsch and Gries 2008). They claim that the results of their studies based on different corpora are very similar. They also argue against predetermined universal lects, such as the spoken – written language distinction. Still, there exists a large body of evidence that shows that even

[1] These studies also demonstrate a complex interaction of processing and conceptual factors, which still requires a clear cognitive interpretation.

unsophisticated *a priori* distinctions be influential predictors of language use. The most powerful source of variation is arguably geographic varieties. For instance, Grondelaers et al. (2002), Bresnan and Hey (2008), Speelman and Geeraerts (2009) and Bresnan and Ford (2010) find that the effects of some semantic parameters on the choice between near-synonymous constructions vary geographically. Stylistic variation, which is associated with registers, channels, and other situational properties, is less frequently explored and usually found to be less outspoken, although there are still interesting effects observed (Grondelaers et al. 2002; Bresnan et al. 2007; Glynn 2007; Speelman and Geeraerts 2009). The presence or absence of lectal variation remains thus an empirical question in every particular case.

1.3. Empirical corpus-based methods of modelling semantics

1.3.1. An overview of existing methods in Cognitive Semantics

Nowadays Cognitive Semanticists use a variety of methods (see Figure 1.3). Most of them employ introspection (Wierzbicka 1985; Talmy 2000). This means that they rely on their own linguistic intuition in describing and explaining linguistic phenomena. This method is closely related to hermeneutics and the idealistic philosophical tradition. The main criticisms levelled at this approach are as follows: it has been shown that speakers' judgments about their own linguistic behaviour are unreliable; no individual is in perfect command of the entire language; the intersubjective comparison of intuitive interpretations is problematic, and so is the full access to meaning as an internal experience (see Geeraerts 2010b). This is not to say that speakers' intuitions are always false: on the contrary, they are often found to converge with the results of empirical studies (see examples in Divjak 2010b: 138; Hilpert 2010). The latter actually seldom completely refute the results of a thorough intuition-based study, mainly adding new details (Arppe et al. 2010: 21). However, introspection becomes problematic if the researcher wants to explore language use outside his or her own linguistic community. In addition, the degree of accessibility of various semantic phenomena to introspection varies (Talmy 2007), and the relationships between introspection and actual linguistic behaviour and comprehension require further investigation. There is another, epistemological, concern: quantifiable analyses are scientifically preferable to qualitative ones because the former are more testable (Popper 1968[1934]: 126), but the results of an introspective analysis are not (easily) quantifiable.

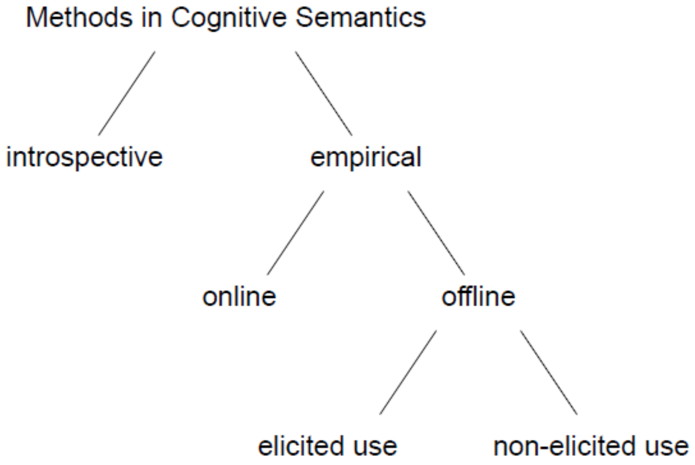


Figure 1.3. Methods in contemporary Cognitive Semantics.

The other approach is empirical and more objective. It allows for repeatability and quantification of findings and, consequently, falsification of hypotheses. These opportunities make this approach more scientific. It is also clear that the gradable salience effects and non-categorical lectal variation can only be explored in a quantitative empirical study (cf. Geeraerts 2005) – the way it has been done in psychology and variational linguistics, respectively.

Empirical methods in Cognitive Semantics display great variation, depending on how meaning is understood and operationalized (cf. Stefanowitsch 2010). Some methods aim at modelling cognitive processes and mental representations associated with semantic phenomena directly online. This approach to semantics, arguably the most embodied one, uses experimental psycho- and neurolinguistic methods, from sorting tasks to priming and neuro-imaging techniques. An example is semantic simulation (Bergen 2007), which studies mental imagery evoked by linguistic expressions. In addition to the semantic contents (e.g. what kind of imagery is evoked), this approach also provides information about *how* words and constructions are processed.

Although experimental evidence is probably the closest to the scientific ideal, this does not mean that all linguists should restrict themselves to experimental study of language-related cognitive processes (cf. Stefanowitsch 2010: 361–365). First, it is difficult to find adequate stimuli to model processing of abstract semantics, like that of syntactic constructions. Second, as was shown in the previous section, meaning is a social phenomenon, and natural communication is difficult to imitate in experimental settings. Third, semantics of some words and constructions can be extremely diverse. As a consequence, the analysis may require a lot of data, which makes experiments a practical challenge. These are probably the main reasons why meaning is more frequently studied offline.

Offline approaches, in their turn, can be based on elicited and non-elicited use. In the first case, they involve elicited data from native speakers (consultants), who are asked to fill in questionnaires. If two different stimuli presented to speakers of one or several languages trigger the use of the same word or construction, then these stimuli are conceptually similar. The approach has been used, in particular, to arrive at typological semantic maps or cluster representations such as in Levinson and Meira (2003), Majid et al. (2007), and Croft and Poole (2008), and to make conclusions about the organization of universal and language-specific semantic spaces of different conceptual domains. Though it gives an idea about the organization of semantic space and categories, this method seems to be less helpful in providing information about semantic variation of individual units.

The other offline approach is based on non-elicited use of language, mainly found in corpora, received from language use *in vivo*, or – less frequently – as a result of an experiment. A detailed description of corpus-based methods in Cognitive Semantics is provided in the next subsection.

This classification is cross-cut by another important distinction, which concerns directness or indirectness of semantic evidence. Direct approaches try to describe the relationships between referents, experience, conceptualizations and linguistic forms by observing immediate links between them. The most direct way of doing so would be by studying the conceptualizations triggered by a word, or the word chosen to represent some conceptual contents. Unfortunately, no one can fully access the conceptual structures directly: even introspection, which might seem to provide the most direct access to meaning, cannot give a pure representation of meaning 'as is' due to the above-mentioned limitations, and also because the meaning has to be represented with the help of some metalanguage (see Geeraerts 2010b), which in turn should be defined with the help of another metalanguage, and so on. A rather direct way is represented by the above-mentioned questionnaire method, where the speakers verbalize the stimuli. As an alternative, one can use non-elicited data with non-verbal stimuli and results of their categorization. An example is the study of Dutch clothing terms in Geeraerts et al. (1994), where the meaning of the lexemes was established with the help of a multimodal corpus with names and pictures of clothing items from fashion magazines. However, this solution is more suitable for concrete lexemes than for abstract categories.

Indirect methods establish relationships between constructions or their senses with the help of a *tertium comparationis* – usually some measurable contextual features, as in the distributional approach to semantics. This approach involves non-elicited natural data extracted from large corpora, although it is becoming increasingly popular to corroborate corpus evidence with experimental results (e.g. Arppe and Järvikivi 2007; Divjak and Gries 2008; Dąbrowska 2009; Bresnan and Ford 2010; Gilquin 2010). The idea behind the distributional approach was aptly formulated by Firth: "You shall know a word by the company it keeps" (Firth 1957: 11). In other words, similar distributional contexts of two linguistic forms (or senses of one polysemous word or construction) are evidence of their semantic similarity. In its most traditional form, it involves looking at the collocations of a word. In constructional approaches to grammar, it is also common to derive the meaning of constructions from the semantic properties of slot fillers (e.g. Stefanowitsch and Gries 2003), thus taking into account the 'internal' context. Conversely, one can infer semantic properties of a verb from the constructions ('alternations' or subcategorization frames) where it can appear (Schulte im Walde 2000). Distributional contexts can also be represented by abstract semantic, syntactic, morphological, prosodic and other features that co-occur with the linguistic unit in question. Many distributional models are believed to represent, at least to some extent, cognitive reality, be it the distributional memory of a speaker (see Baroni and Lensi 2010), probabilistic knowledge of language (e.g. Bresnan and Ford 2010), or mental organization of lexicon (Divjak and Gries 2008). However, distributional models may be difficult to interpret in a theoretically meaningful way due to their indirectness and the presence of all kinds of contextual 'noise'. More about these methods follows in the next subsection.

The indirect distributional approach has also been used in studies outside corpus linguistics. For example, Levin's (1993) analysis of verb classes is based on introspective observations of the use of verbs in specific grammatical alternations (cf. Schulte im Walde 2000, who applies a similar approach but uses evidence from a large corpus). In online experimentation and elicitation tasks, one can use distributional evidence, too, when studying collocational patterns and semantic preferences in elicited sentences produced by subjects (e.g. Dąbrowska 2009).

Unfortunately, there seems to be a negative correlation between directness (interpretability) and objectivity of evidence in semantics. Introspection is the most direct but subjective, whereas corpus-based methods are objective but their results are more difficult to interpret. This might be one of the factors that has hindered the development of empirical methods in Cognitive Linguistics despite its acceptance of the usage-based paradigm.

1.3.2. Corpus-based distributional methods

Distributional corpus-driven approaches are arguably the most popular in empirical Cognitive Semantics, except for semantically oriented typology, where data may be sparse, and figurative language studies, which often focus on the way metaphors and metonymies are processed. There are many different approaches, which differ in the design of various stages of analysis. Below is a brief overview.

Stage 1: collecting the data. This process can be automated or manual. In the former case, researchers use large-scale parsed and tagged corpora. An example is the computationally intensive Vector Space models of lexical semantics, used for modelling polysemy or various semantic relationships between words (word forms). The relationships between the objects of study are established by comparing their distributional contexts defined in a variety of ways (as documents, 'bags of words' on the right and left from the word, syntactic subcategorization frames, etc.) with the help of a co-occurrence measurement, such as the Pointwise Mutual Information index. Originating in Computational Linguistics (Lin 1998), Vector Space Models are only finding their way to lexical Cognitive Semantics at the present (e.g. Heylen et al. 2008). These models are extremely sensitive to topic-related contextual differences and semantic prosody, which can be seen as both their power and limitation. For instance, Peirsman et al. (2010) demonstrate the change in lexical associations of the word *Islam* after the terrorist attacks of 9/11 in Dutch newspapers. The method also allows the researcher to identify lectal variants of similar concepts, as Peirsman (2010) shows. Application of the method for constructional semantics is less straightforward, although there have been a few attempts. For example, in what they call a radically data-driven version of Construction Grammar, Levshina and Heylen (In preparation) use fully bottom-up semantic classes of constructional slot fillers to model the conceptual differences between near-synonymous constructions.

Another (relatively) unsupervised approach is Collostructional Analysis developed by Stefan Gries and Anatol Stefanowitsch (Stefanowitsch and Gries 2003; Gries and Stefanowitsch 2004; Wulff et al. 2007, etc.). It is a family of methods based on measuring attraction/repulsion between constructions and collexemes (lexical slot fillers). Collostructional Analysis uses Fisher's exact test for every collexeme found in the data to identify significantly attracted or repelled collexemes on the basis of several frequency measurements. A subtype of

Collostructional Analysis, Distinctive Collexeme Analysis, identifies the collexemes that are more distinctive of one construction (or its variant in one lect) than of another one (or the counterpart in another lect).

Semantic Vector Spaces and Collostructional Analysis are arguably the most objective methods among the distributional techniques. On the other hand, they have to combat with problems such as polysemy of lemmata, tagging deficiencies and other noise that can influence the results.

Other approaches are based on manual coding of observations for a number of semantic, syntactic and other variables and provide better control of the results. For example, the Behavioural Profiles approach, which is based on the ideas of Atkins (1987) and Hanks (1996) and recently elaborated by Divjak and Gries (e.g. Divjak 2006; Gries 2006; Divjak 2010b), allows the modelling of both polysemy and near-synonymy. A disadvantage of this method when applied to polysemy is that it requires pre-defined senses from a dictionary or another source.

Stage 2: exploring the data. Next, the co-occurrence data found in Semantic Vector Spaces or Behavioural Profiles is represented as vectors of co-occurrence values (Semantic Vectors), or as average profiles with percentages of each level for every variable (Behavioural Profiles). A range of standard and novel multivariable techniques are available for exploring the relationships between these vectors. In practice, however, the semantic similarity between every two senses or words in Semantic Vector Spaces is operationalized as the cosine between the corresponding vectors. The higher the cosine value, the larger the similarity. In Behavioural Profiles, the senses (near-synonyms) are clustered according to their profiles. By zooming in on the profiles, one can find the features that are the most distinctive of a specific sense or a near-synonym.

The output of a Collostructional Analysis is usually a list of collexemes with their measures of attraction or repulsion to or from the construction in question. These lists should next be interpreted semantically, which is not an easy task (cf. Gries and Stefanowitsch 2010). The method has been recently criticized by Bybee (2010) and Schmid (2010) for several conceptual problems. Most importantly, it does not take into account the semantic similarity between the collexemes. As a result, the role of the low-frequency collexemes that are semantically similar to high-frequency collexemes and thus have a high degree of typicality, is ignored.

Two other techniques involve low-dimensional representations of the conceptual space. (Multiple) Correspondence Analysis in semantic research was introduced by Glynn (2007), who models both synonymy and polysemy of lexemes with the help of their semantic features. Correspondence analysis is based on chi-squared distances between categorical variables, which are visualized on low-dimensional maps. A more sophisticated version has been presented in Levshina et al. (In press), where exemplar representations were added, which allows one to investigate the structure of categories and their semantic overlap more easily. The new approach also makes a sharper distinction between form and function, so that the contextual variables serve as a basis for a semantic map with interpretable semantic dimensions. Yet, the technique is sensitive to low-frequency semantic features, and the chi-squared distance between exemplars or semantic features is not always easy to interpret.

An innovative way of visualizing diachronic changes in semantics has been recently suggested by Martin Hilpert (Submitted). In essence, the data are very similar to Behavioural Profiles, but visualization is done with the help of a set of Multidimensional Scaling maps, which correspond to different historical periods. The latter are then compiled in a kind of a flip book, which shows the semantic development of constructions over time.

Stage 3: confirmatory tests. After exploring the data and thinking about possible hypotheses, it has become almost standard practice to see if the results can be extrapolated to other data (ideally, the entire language) by applying statistical significance tests. Although there is a range of hypothesis-testing statistical techniques (*t*-tests, chi-squared tests, etc.), the most popular method is arguably multiple (logistic) regression analysis because it allows for the testing of all variables of interest simultaneously and modelling their individual impact on the outcome – the speaker's choice between two or more lexemes or constructions. The technique thus reveals the distinctive semantic features of the linguistic units. It can also be applied to several senses of one word (e.g. the adjective *awesome* in Robinson 2010) to arrive at their distinctive properties. Multiple regression analysis allows the researcher to integrate cognitive and social factors in one model (e.g. Bresnan et al. 2007; Speelman and Geeraerts 2009) or compare the relative impact of cognitive variables in several lectal models (e.g. Grondelaers et al. 2002; Szmrecsanyi 2010). Random effects are sometimes added to the model to filter out the 'noise' caused by lexical effects or individual differences between speakers or corpus sources, and test the significance of high-level generalizations (e.g. Bresnan et al. 2007; Divjak 2010a; Levshina et al. Submitted). Some of the limitations of regression analysis are the restricted number of predictors that can be tested simultaneously in a model without running the risk of overfitting, and problems with interpretation of multiway interactions between the variables.

In this study I propose a new method of modelling semantic and lectal variation both in the structure of one construction and between two or more semantically related units. The approach is based on semantic similarity between constructional exemplars (unique observations), although in principle it can be applied to lexical variation, too. It is truly bottom-up, so that generalizations in the form of semantic dimensions and usage clusters (or lower-level schemata) emerge from a sample of constructional exemplars. The approach involves a range of visualization techniques and statistical tests that enable a researcher to give a semantic interpretation of the distributional patterns and disentangle semantic and lectal sources of variation.

1.4. Summary

This chapter has outlined the theoretical and methodological framework of this study – empirical lectally enriched Cognitive Semantics, which belongs to the family of usage-based approaches to language. This is an interdisciplinary field, which poses fuzzy boundaries between lexicon and grammar. I have shown that, if one takes the cognitive commitment of Cognitive Linguistics seriously, it is necessary to take into account different salience effects, as well as the intra- and intercategorial perspectives on semantic structure. As recent studies in Cognitive Sociolinguistics demonstrate, it is also desirable to test if there is significant lectal variation in the use of linguistic units.

Methodologically, this study belongs to the family of empirical distributional methods based on corpus data – the currently prevailing approach in empirical constructional semantics. Because the theoretical goals outlined above are difficult to achieve with the current quantitative methods, this study presents an innovative approach to modelling semantic and lectal variation. The approach is applied to the Dutch causatives with *doen* and *laten*, which are introduced in the next chapter.

Chapter 2. Variation of the Dutch causative constructions with *doen* and *laten*

The Dutch causative constructions with *doen* and *laten* have received a significant amount of attention in linguistics. Most of the recent studies treat the constructions as independent symbolic units similar to simple clause structures. This approach is fundamentally different from the generative tradition, which assumes that analytic causatives are derivations from two independent clauses (see Kemmer and Verhagen 1994 for more details). The conceptual difference between the near-synonyms has been explored in a few corpus-based studies (Kemmer and Verhagen 1994; Verhagen and Kemmer 1997; Degand 2001; Stukker 2005), where it is interpreted as the difference between direct and indirect causation. At the same time, Verhagen and Kemmer (1997) and Stukker (2005) proposed a semasiological analysis of each construction. In the onomasiological multivariate analysis in Speelman and Geeraerts (2009), more factors were taken into account, and lectal and collocational dimensions were added to the model. In this chapter, I discuss the above-mentioned contributions, as well as studies of some specific subconstructions with *doen* and *laten*. A separate section is devoted to synchronic and diachronic variation in the use of the constructions and to several hypotheses that try to interpret this variation.

2.1. Form and function of Dutch causative constructions with *doen* and *laten*

The Dutch causative constructions with *doen* and *laten* consist of the

Auxiliary Predicate (*doen* “do” or *laten* “let”), the Effected Predicate and several nominal slots, as shown in the example (1). The Causer (here, the police) is the initiator of the event, the Causee (the car) is the entity that performs the action or undergoes a state specified by the Effected Predicate. The causation involves two events: the causing event (the police did something to make the car stop) and the caused event (the car stopped).

(1)	*De politie*	*deed/liet*	*de auto*	*stoppen*.
	The police	did/let.PAST	the car	stop.INF
	Causer	Aux. Predicate	Causee	Effected Predicate

“The police stopped the car (let the car stop).”

Analytic causatives, which leave the causing event denoted by the auxiliary unspecified, occupy an intermediate position between lexical causatives (e.g. *break*, *kill*), where the causing and caused events are maximally integrated and indistinguishable, and clausal structures with causal connectives (e.g. *because*, *therefore*), which denote maximally distinct and specified causing and caused events (e.g. Stukker 2005; Shibatani and Pardeshi 2002).

In the cases of transitive Effected Predicates, there can be a third participant, the Affectee – the energy 'sink' and the end of the causation chain.[1] In (2), this role is played by the city:

(2)	*De generaal*	*liet*	*het leger*	*de stad*	*vernielen.*
	The general	let	the army	the city	destroy
	Causer		Causee	Affectee	

“The general ordered the army to destroy the city.”

[1] Degand (2001: 181) and Stukker (2005: 43–44) assign the role of the Affectee to all kinds of objects of the Effected Predicate: direct, indirect and prepositional objects, including subordinate clauses. Those are interpreted as entities that are literally or metaphorically affected by the caused event. The present study follows this approach.

In addition to these roles, Draye (1998) suggested a fourth nominal participant, the so-called Interestee, which corresponds to the indirect object of ditransitive Effected Predicates. Such contexts, however, are very rare.[1]

The semantic difference between *doen* and *laten* has been examined in a number of corpus-based studies (Kemmer and Verhagen 1994; Verhagen and Kemmer 1997; Stukker 2005). It has been suggested that the difference lies in the construal of the situation as direct or indirect causation. The use of *doen* shows that the situation is construed as direct causation, i.e. "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). Leuwenthal writes,

> "Because the action of the causer is seen as a sufficient condition to realize the effect, the causee's role can be seen as minimal and not relevant in the realization of the effect, although the causee is in fact the one that "carries out" the activity. There is no intention needed from the side of the causee to carry out the effect, the effect happens beyond his consciousness or control." (Leuwenthal 2003: 101)

Indirect causation emerges when "it is recognized that some other force besides the initiator is the most immediate source of energy in the effected event" (Verhagen and Kemmer 1997: 67). This immediate source of energy is either a volitional animate Causee, as in (3a), or an external force like gravity working upon the Causee, as in (3b). Although the Causer is still the main entity responsible for the causal event, his or her energy is "not a sufficient condition for the realization of the effect"

[1] Draye also argues that the participants that can be interpreted as semantic subjects of the Effected Predicates expressed by verbs of perception, are not Causees, but Interestees. In this study, however, the participants are treated according to their surface forms and positions, not deep thematic roles.

(Loewenthal 2003: 101).

(3) a *De trainer liet zijn spelers loopoefeningen doen.*
the coach let his players run-exercises do
"The coach had his players do running exercises."

b *Hij liet het water weglopen.*
he let the water away-run
"He let the water drain out."

If we revisit one of the above-mentioned examples, which is presented for convenience as (4), the following interpretation is possible. The choice of *deed* (the past form of *doen*) would activate a situation when the police caused the car to stop regardless of the driver's intentions, e.g. by blocking the road, whereas *liet* (the past form of *laten*) would suggest that the police signalled the car to stop, and the driver stopped the car consciously.

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

In addition, *laten* allows for a permissive reading, which can be seen as the extreme case of indirectness. In Talmy's force dynamics (Talmy 2000: Ch. 7), causation *per se* (coercion, impingement) means that the stronger Causer (the antagonist, in Talmy's terms) overrides the intrinsic tendency of the Causee (the agonist) towards rest or motion. Letting (enablement or permission) involves a Causer who fails, deliberately or not, to override the Causee's intrinsic tendency. In fact, constructions with *laten* range from enabling/permissive to coercive meanings, with a number of ambiguous cases in between. Compare the letting context in (5a) with the ambiguous (5b) and coercive one (5c):

(5) a *De politie liet de dader ontsnappen.*
the police let the criminal escape
"The police let the criminal escape."

b *Hij liet iedereen zijn roman lezen.*
he let everybody his novel read
"He had/let everyone read his novel."

c *De trainer liet de spelers loopoefeningen doen.*
the coach let the players run-exercises do
"The coach had the players do running exercises."

As Verhagen and Kemmer (1997) demonstrate, directness and indirectness of causation are closely associated with the configurations of the semantic classes of the Causer and the Causee. Thus, if both the Causer and the Causee are animate (human), one can expect the causation to be indirect because a human being cannot affect another mind directly, telepathy disregarded (Verhagen and Kemmer 1997: 71). This type of causation is often labelled as inducive causation, e.g. (6a). In contrast, physical entities normally affect other physical entities directly, as in (6b). They can also affect a human mind directly as a cognitive stimulus does, e.g. (6c). This causation type is called affective causation. In the case of volitional causation, with an animate causer and a non-mental Causee, no predictions can be made because a human being can change the world both directly and indirectly, e.g. with the help of automation, as in (6d).

(6) a *De trainer liet de spelers loopoefeningen doen.*
the coach let the players run-exercises do
"The coach had the players do running exercises." [inducive]

b *De aardbeving deed de muren trillen.*
the earthquake did the walls shake
"The earthquake made the walls shake." [physical]

c *Je kapsel doet me denken aan een vogelnest.*
your hairstyle does me think to a bird-nest
"Your hairstyle reminds me of a bird's nest." [affective]

d *De machinist liet de motoren draaien.*
the engine-driver let the engines run
"The engine driver had/let/left the engines run(running)." [volitional]

These regularities are not absolute. For example, (7) demonstrates that affective causation can also be triggered by a human Causer. In this case neither the Causer nor the Causee act as volitional agents. It is something in the psychiatrist's appearance or behaviour that made the Causee think of his/her mother (Verhagen and Kemmer 1997).

(7) *De psychiater deed me aan mijn moeder denken.*
the psychiatrist did me to my mother think
"The psychiatrist made me think of (reminded me of) my mother." (Verhagen and Kemmer 1997: 73)

Using *laten* in this context would imply that the psychiatrist asked the patient to think about his or her mother. This would be a case of inducive and less direct causation.

(8) *De psychiater liet me aan mijn moeder denken.*
the psychiatrist let me to my mother think
"The psychiatrist had/made me think of/about my mother."
(Verhagen and Kemmer 1997: 73)

This has important methodological consequences: the properties of the individual participants can be seen only as indirect evidence of the constructional meaning, as more or less typical situations that might be compatible with the construal (see the discussion of direct and indirect semantic evidence in the previous chapter, Section 1.3). Even volitionality of the Causee, whose role seems to be crucial in the construal, does not guarantee a perfect distinction between *doen* and *laten* (Degand 2001: Ch. 6).

Nevertheless, these tendencies are quite important. Treating *doen* and *laten* as prototypical categories, Stukker (2005: 63–67) shows that physical and affective causation together constitute the semasiological prototype of *doen*,[1] whereas inducive causation is typical of *laten* (the prototype is defined conceptually: it is the sense that serves as the starting point for extensions). Other configurations are extensions from these prototypical senses. The semasiological prototypes in Stukker's study coincide with the onomasiological prototypes. Stukker also notes that *doen* is more frequently used in untypical situations than *laten*, "construing the causal relation in a non-standard way" for rhetorical purposes (Stukker 2005: 67). This asymmetry may be due to the quantitative and qualitative dominance of *laten* as the default causative, which is acceptable in most cases, and the markedness of the functionally and quantitatively restricted *doen* (Speelman and Geeraerts 2009, see also Section 2.3).

[1] Speelman and Geeraerts (2009) propose a more specific hypothesis – namely, that *doen* is restricted to direct physical causation only, – but they do not test this hypothesis in their study.

All these studies focus on general semantic features of *doen* and *laten* and some of their top-level subconstructions. However, as Speelman and Geeraerts (2009) show, there are additional effects that influence the use of the auxiliaries. These are fixed lexical expressions, such as *doen denken aan* “remind of” or *laten zien* “show”. Speelman and Geeraerts demonstrate that the lexical attraction between the verbs that fill the Effected Predicate slot and the auxiliary (operationalized with the help of Collostructional Analysis) is stronger in the case of *doen* than in the case of *laten*. They interpret it as evidence of the marginal status of *doen* in contemporary Dutch. From the purely conceptual perspective, lexical fixation can be seen as a symptom of extremely high conceptual integration of the causing and caused events, and therefore a sign of very direct causation (cf. Duinhoven 1994a, 1994b). The existence of such collocations alongside the generalizations is in line with the non-reductionist nature of the speaker's knowledge of constructions, which brings us to the topic of the next section.

2.2. The constructional network of *doen* and *laten*

As noted in the previous chapter (Section 1.1), one of the distinctive features of constructional semantics is that the meaning of more schematic constructions is contributed by the semantics of more specific form-meaning pairings. The latter behave like hyponyms, covering a part of the semantic space of the superordinate construction. In fact, the causation frames involving animate or inanimate Causer and Causee discussed above can be regarded as very schematic prototypically organized subcategories of *doen* and *laten* with a specific meaning (affective, physical, inducive, volitional causation). The previous studies also discussed some other lower-level schemata of *doen* and *laten*. Below I present an overview of their main findings, focusing on the specific contributions of the subschemata to the semantics of their 'parents'.

2.2.1. Transitive and intransitive constructions

Kemmer and Verhagen (1994) suggest a cross-linguistic typology of causative constructions, distinguishing between intransitive (two-participant) and transitive (three-participant) constructions. They claim that intransitive causative constructions (e.g. *He made me laugh*) are extensions of transitive constructions (with such predicates as *break*, *kill*, *eat*, etc.), whereas transitive causatives (e.g. *The government made people pay extra taxes*) are extensions of ditransitive clauses (for instance, those with the verbs *give*, *send*, *show*). By extension they understand “an asymmetric synchronic relation between two structures (…) viewed as one of inheritance of particular properties from the more basic one by the more complex one” (Kemmer and Verhagen 1994: 128). This means that intransitive causative constructions share the semantic prototype with simple transitives: an individuated agent exerts energy on an individuated affected patient (Kemmer and Verhagen 1994: 126–127). In intransitive causative constructions in many languages the first participant is marked as the subject, and the second participant is usually marked as the direct object. However, in intransitive causative constructions the causation is less direct than in simple transitive clauses, and the Causee has some degree of autonomy (see more details in Kemmer and Verhagen 1994: 127–128). As for transitive causative constructions, their middle participants, the Causees, are similar to some extent to dative case arguments in ditransitive clauses, being less crucial to the structure of the event than the other two participants. Another common role of the Causee in transitive causative constructions is that of a metaphorical instrument. Because of this variation, transitive causative constructions, like simple ditransitives, display more cross-linguistic variation in marking of the Causee than intransitive causative constructions (see the next subsection).

It is not by chance that, according to all previous corpus-based studies, transitive causative constructions with *laten* occur more frequently than the ones with *doen*. Indirectness of causation combines well with peripherality of Causees, which is typical of transitive causative constructions.

2.2.2. Marking of the Causee

As mentioned above, there are several ways of expressing the Causee in Dutch (according to Verhagen 2007, this is evidence of tight integration of the causatives with the simple clause constructions). In addition to the 'default' zero-marked nominal phrase, the Causee can be expressed by a prepositional phrase with *aan* “to” and *door* “through”, especially in transitive causative constructions with *laten*. The marking depends on the semantic role of the Causee. *Door* marks agentive Causees (note that the same preposition is used to mark the agent or force in the passive construction), whereas *aan* marks recipient-like Causees (e.g. Dik 1980: 67). The use of prepositions also means a lesser topicality/affectedness than in the case of a zero-marked nominal phrase. In fact, the degree of topicality/affectedness corresponds to the following cline (Kemmer and Verhagen 1994: 133–134):[1]

zero-marked NP > *aan* + NP > *door* + NP

Compare the most patient-like, maximally topical and affected Causee expressed by a zero-marked NP *de spelers* in (9a), the medium-affected and topical Causee-recipient *aan iedereen* in (9b), and the minimally affected and maximally peripheral Causee-agent *door een architect* in (9c):[2]

[1] Note that a zero-marked NP is maximally affected only in the cases when the construction conveys coercion. If the causative expresses letting, as in (5a), the Causee, which is normally zero-marked, is not affected at all.

[2] In this respect, *laten* is similar to its German cognate *lassen*, which allows for marking of the Causee with *von* “by” in agentive causative events (Draye 1998). Similarly, in the French causative

(9) a *De trainer liet de spelers loopoefeningen doen.*
the coach let the players run-exercises do
"The coach had the players do running exercises."

b *Hij liet de brief aan iedereen zien.*
he let the letter to everybody see
"He showed the letter to everyone."

c *Ik liet mijn huis ontwerpen door een architect.*
I let my house design by an architect
"I had my house designed by an architect."

Note that the possibilities of prepositional marking are limited. *Aan* occurs almost exclusively with verbs of perception, whereas *door* is used normally with transitive Effected Predicates. Only one verb, *lezen* "read", combines with zero-marked NPs, *door* and *aan* (Dik 1980: 56).

It is clear that direct causation is not compatible with peripheral and autonomous Causees. Therefore, *doen*, which denotes direct causation, disfavours prepositional marking (Kemmer and Verhagen 1994: 144), although not categorically – see, for instance, example (5) in Chapter 6.

2.2.3. Causeeless constructions

Causeeless constructions with *laten* are discussed in detail in Loewenthal (2003). As a rule, these constructions contain a transitive Effected Predicate and an explicit Affectee. An example is (10):

construction with *faire* "make", the preposition *par* "by" marks an agentive Causee, whereas *à* "to" shows that the Causee is more affected as a beneficiary or recipient (Degand 1996). The English analytic causatives with *make*, *have* and *get* allow for prepositional marking of the Causee (*by* or sometimes *to*) only when the Effected Predicate is in the form of the past participle (Levshina et al. In press).

(10) *Ik liet mijn huis schilderen.*
I let my house paint
"I had my house painted."

The interpretation of the Causee is highly schematic: it is "exhausted by the information provided by the effected predicate" (Verhagen and Kemmer 1997: 63) and reduced to the role of an abstract 'painter'. The Causee is not expressed in this example because the focus is on the change of the Affectee and it is not particularly important who exactly carried out the caused event (cf. agentless passive constructions). Many examples of causeeless constructions evoke the service frame, as in (10), institutional causation and other types of social interaction.

Another kind of causeeless constructions with *laten* has permissive semantics and the coreferential Causer and Affectee. This construction commonly conveys some negative influence that the Causer manages or fails to prevent, as in (11):

(11) *Hij liet zich niet misleiden.*
he let himself not mislead
"He didn't let himself be mislead."

The reflexive construction, especially with inanimate Causers, sometimes expresses middle voice events. In such cases, the Causer, which is also the semantic patient, facilitates or hinders the energy flow due to its inherent properties (see more in Davidse and Heyvaert 2003). For example,

(12) *Cultuur laat zich niet makkelijk exporteren.*
culture lets itself not easily export
"Culture cannot be exported easily."

Another common type of causeeless constructions with *laten* involves verbs of perception (*zien* "see", *horen* "hear") and *weten* "know", as in (13).

(13) *De minister heeft laten weten dat hij ontslag neemt.*
the minister has let know that he resignation takes
"The minister has made known that he is going to resign."

In these constructions the role of the Causee is that of a recipient, whose identity is either unclear or can be inferred from the context. The constructions are frequently used in situations that involve official communication and mass-media, and the addressee is the general public, as in (13).

To summarize, causeeless constructions with *laten* can express mild coercion such as in the service frame, middle voice events, or providing information with an unidentified addressee. Note that *doen* also allows for implicit Causees. Such structures are very common in the contexts that denote affective causation:

(14) *Deze film doet denken aan Fellini.*
this film does think to Fellini
"This film reminds of Fellini."

It seems that the implicitness of the Causee in this intransitive construction leads to a closer integration of the causing and caused events because of the lack of any other participants between them. Compare this construction with the similar *faire* + Verb structure in French (Achard 2002):

(15) *Marie a fait pleurer sa soeur*
Marie has done cry her sister

“Marie has driven her sister to tears.”

However, the causeeless constructions with *doen* perform an additional pragmatic function. The Causee is omitted to show that the result of causation is inevitable, whoever the cognizer might be. In this way, the speaker focuses on the objective properties of the stimulus denoted by the Causer. Note that a similar pragmatic shift takes place with the causeeless *laten weten*, *laten zien* and some other constructions. The speaker focuses on the act of saying or showing, regardless of the fact whether the addressee obtains the information or not.

***2.2.4. Lexically specific constructions with* doen *and* laten**

Besides *doen denken aan*, there are plenty of frequent expressions with *doen*, e.g. *iemand iets doen opmerken* “draw someone's attention to something”, *iemand naar iets doen verlangen* “make someone crave for something”, *van zich doen spreken* “make people talk about oneself, make one's mark”, *iemand iets doen toekomen* “send someone something” (often about official documents), *hoop doet leven* “while there is life there is hope (lit. hope makes live)”. There are some metaphorical expressions, e.g. *een belletje doen rinkelen* “ring a bell, be familiar” and *het tij doen keren* “turn the tide”. The constructionist approach to language does not make qualitative distinctions between these expressions and non-idiomatic constructions (Kemmer and Verhagen 1994: 147).

Although *laten* overall seems to be more productive than *doen* (Speelman and Geeraerts 2009), the former is used in a large number of fixed and idiomatic combinations, too. For example, *laten zien* “show” and *laten weten* “inform” with explicit or implicit Causees (see the previous section), refer to transfer of information. Note that in other languages these meanings are often expressed lexically (cf. English *show*, German *zeigen*, Swedish *visa*, French *montrer*, Russian *показывать*), although periphrastic expressions are often acceptable, too. The lexical alternatives are also available in Dutch, for example *tonen* “show”, *informeren* and *berichten* “inform”.

There are also many idiomatic figurative expressions with *laten*, most of them with the meaning “abandon” and intransitive Effected Predicates, for example *iemand laten vallen* “ditch someone (lit. let someone fall)”, *een plan*, *alle hoop laten varen* “abandon a plan, all hope (lit. let a plan, all hope go)”, *iemand in de kou laten staan* “leave someone out in the cold, abandon in a difficult situation”, *links laten liggen* “ignore, shrug off (lit. let lie on the left)”. Other expressions are about losing control over the situation and/or being manipulated: *zich laten gaan* “let oneself go”, *zich van de wijs laten brengen* “lose one's head (let. let bring oneself off the tune)”, *zich laten doen* “let people push oneself around (lit. let do oneself)”. For more examples, see Coopmans and Everaert (1988). There are also expressions that contain rare infinitiveseseses that are used only in set expressions with *laten*, e.g. *laten betijen* “leave alone, neglect” and *het laten afweten* “fail, not show up” (see Coopmans and Everaert 1988: 92–93). In Schmid's (2010) terminology of relationships between a construction and its slot fillers, these verbs have 100% reliance on *laten*. Interestingly, such combinations are difficult to find with *doen*. Apparently, *laten* is a semantically stronger construction, which lets the other verbs fully rely on its semantics, whereas *doen* probably itself needs semantic support from the other slot fillers.

There are quite few frequent Effected Predicates that are used interchangeably with *doen* and *laten.* One of them is *geloven* “believe”. According to the folk model of the mind, believing is a mental process that is usually controlled by the believer (D'Andrade 1987). Therefore, *geloven* should occur mostly with *laten*. Yet, it is frequently encountered with *doen*. It appears that there is a subtle semantic difference between the two constructions. Unlike *laten geloven*, the variant with *doen* normally refers to the Causer's unsuccessful attempts to make the Causee believe something (cf. Verhagen and Kemmer 1997: 75). These attempts are identified by the Causee and resisted, so the causation does not take place. Compare (16a) and (16b):

(16) a *Het is niet zo slecht als men ons wil doen*
it is not so bad as they us want do
geloven.
believe
"It is not so bad as they want to make us believe."

b *Tot wanneer mag je kinderen laten geloven*
Till when may you children let believe
in de Kerstman?
in the Christmas-man?
"Until what age should you let children believe in Santa Claus?"

Thus, both *doen* and *laten* occur in a range of specific formal patterns associated with transitivity, Causee marking, reflexivity, etc., which differ semantically in line with the directness/indirectness distinction, although at the level of lexically specific subconstructions this difference can be a matter of a very subtle construal. From the methodological point of view, this suggests that many fine-grained semantic and formal features should be taken into account when modelling the semantics of the constructions.

2.3. Dutch causative constructions: diachronic and synchronic variation

The first occurrences of the causative *doen* are found in texts from the 13th century – the earliest period for which sufficient linguistic evidence is available (van der Horst 1998). According to one hypothesis, the causative *doen* emerged in that period in order to compensate for the decrease in productivity of morphological causatives of the type *drink* "drink" – *drenken* "give water (to)" (van der Horst 2008: 225). In the early period, *doen* referred to different types of causation, which can be interpreted (at least by contemporary speakers) as direct and indirect. The Early Middle Dutch dictionary (Pijnenburg 1997) names two main senses: bringing about a physical or mental caused event, as in (17), and the situations when someone was assigned, requested or ordered to do something, e.g. (18):

(17) *Die lettre doedt. die gheest doet leuen.*
the letter kills the spirit does live
"The letter killeth, the spirit giveth life. "

(18) *wi daden seggen, dat si doot ware*
we did say that they dead were
"We ordered (someone) to say that they were dead" (Duinhoven 1994a: 112).

Interestingly, quite a few collocations are reported from the early periods with the mental verbs that are now more commonly used with *laten*: *doen (te) weten* "inform" and *doen (te) verstaen* "make understand" (van der Horst 2008: 426).

The first attestations of the auxiliary *laten* in Dutch are also found in the 13th century. It was used to express permission and enablement, as its

Gothic precursor *lētan*, which denoted letting/permission, as in (19). However, some occurrences of indirect coercive causation, as in (20), were also attested in that period.[1] This semantic extension might have been affected by a functional levelling of *doen* and *laten* in early Middle Dutch (van der Horst 1998).

(19) *lat dise arme kinde leuen*
let these poor children live
"Let these poor children live." (van der Horst 1998: 62)

(20) *eine quitaine laet mi maken*
a quitaine let me make
"Have a quitane[2] be made for me." (van der Horst 1998: 62)

It seems that some conceptual patterns that were formerly associated with *doen*, are now more typical of *laten*. Is this evidence of language change? There are two opposite views. Let us consider them in turn.

The first view was expressed by Duinhoven in his articles based on diachronic evidence (1994a; 1994b), and later supported by a synchronic variational analysis in Speelman and Geeraerts (2009). According to this point of view, *doen* is an "obsolescent form with a tendency towards semantic and lexical specialization" (Speelman and Geeraerts 2009: 200). Duinhoven argues that *doen* has lost a larger part of its semantic repertoire since the early days, when it conveyed any kind of causation with an active Causer. First, the auxiliary lost the indirect causation sense, and is now on the way to losing the direct causation function. Duinhoven claims that this process took place because of the changes in the function of the Dutch infinitive. The latter used to perform an adverbial function, and therefore

[1] Interestingly, its German cognate *lassen* extended its semantics from letting to coercive causation, too (Soares da Silva 2007: 188–191).

[2] quitaine: a medieval instrument used for sward-fighting (van der Horst 1998: 62).

the caused event expressed by the infinitive was semantically autonomous from the causing event expressed by *doen*. However, this function has been lost. As a result, *doen* has become so tightly integrated with the infinitive that the present-day construction can express only extremely direct causation. The other *doen*'s functions (indirect coercive causation and even to some extent direct causation) have been taken over by *laten*. The latter could fill in the gaps because it denotes less tightly integrated events, due to the relatively passive role of the Causer. This point of view corresponds to the second (ii) type of language change (see Chapter 1, Section 1.2), which involves a change in the category structure.

This opinion differs dramatically from Verhagen's (1994a, 1994b, 2000), who believes that the abstract meaning of *doen* (direct causation) has not changed fundamentally, at least since the 18th century, when interpersonal causation events were still commonly categorized with *doen*. What has changed, is the construal of social relationships and authority. The speakers no longer construe social interaction as involving the Causer's full control over the Causee. Thus, the conceptual contents associated with *doen* have not changed, but the conceptualization of certain experiences has altered. Verhagen's arguments are as follows: first, the relative frequency of *laten* has not increased, as one could expect in accordance with Duinhoven's hypothesis. Second, *doen* is still fairly frequent in many registers, including fiction – probably due to the contemporary practice of narration, which involves a highly subjective perspective that can manifest itself in reference to affective causation. Third, *laten* with authoritative Causers has also become less frequent. As Verhagen puts it, "authority has become a less important aspect of our models of interpersonal relations (if not of these relations themselves)" (Verhagen 2000: 274). This process is thus very close to the first type (i) in the above-mentioned typology of sources of language change. In other words, the experience of the speakers has changed, but not the category, although, of course, one can say that the meaning (concept) has changed even if the change concerns only the degree of entrenchment of certain usage patterns.

From the synchronic point of view, there is also substantial regional and functional variation in the use of the two causatives. Speelman and Geeraerts (2009) have shown that *doen* is more favoured in Belgian Dutch and formal prepared speech than in Netherlandic[1] Dutch and informal spontaneous communication. Their conclusions are based on a multivariate model with semantic factors controlled for (cf. ANS 1997: 1015–1017). Speelman and Geeraerts interpret these facts as evidence that *doen* is becoming obsolete. In fact, Belgian Dutch is believed to have retained more archaic features than Netherlandic Dutch due to several sociohistorical reasons, such as belated standardization (Geeraerts et al. 1999: 13–18) and persistence of dialects (Auer 2005: 25). On the other hand, formal registers retain more archaic features than informal speech does.

However, the geographic variation could be interpreted in line with Verhagen's hypothesis, too. One could argue that the higher tolerance of *doen* in Belgian Dutch might be due to the more important role of authority in the culture in comparison with the Netherlands. It is interesting that Hofstede (2001) in his quantitative cross-cultural study finds that Belgium has a higher power distance index than the Netherlands. This index correlates with the perception of authorities as different from common people, which implies that social status is a more salient attribute in Belgium than in the Netherlands.

All the above has important consequences for the present synchronic study of variation in contemporary Dutch. Regardless of the actual cause of the variation, one can expect the semantics of *doen* in Belgian Dutch to be less contextually restricted than in the Netherlandic variety, especially

[1] Here and throughout I use the adjective Netherlandic to refer to the variety of Dutch spoken in the Netherlands.

in the interpersonal causation area. In addition, one can try to disentangle the sources of variation, albeit indirectly. If the Belgian *laten* is less frequent, and less active in the domain of direct causation than its Netherlandic counterpart, this will be evidence in favour of Duinhoven's hypothesis that the semantics of *doen* has been taken over by *laten*.

2.4. Summary

In this chapter I have presented a number of hypotheses and facts about the causative constructions with *doen* and *laten*. Most studies support the direct/indirect causation distinction between the two auxiliaries. This distinction is accompanied by a range of semantic, formal and lectal features shown in a compact form in Table 2.1.

The formal and semantic variation of *doen* and *laten* shows that a variety of minute contextual detail may be helpful in the interpetation (recall, for example, the subtle difference between *doen geloven* and *laten geloven*). In these circumstances, a multivariate analysis is a necessity. This brings us to the next chapter, which describes the quantitative approach employed in the present study.

	doen	***laten***
General meaning	(more) direct causation	(more) indirect causation
Sema- and onomasiological prototypes	affective and physical causation	inducive causation
Transitivity of the Effected Predicate	less frequently transitive	more frequently transitive
Semantic role of Causer	more active	less active
Semantic role of Causee	more patient-like	less patient-like
Marking of Causee	zero-marked	zero-marked, *aan* and *door*
Presence of Causee	more frequently explicit	more frequently implicit
Energy Flow	always from Causer to Causee	Causer can be affected
Lexical fixation (wrt. Effected Predicates)	stronger	weaker (but high reliance of some Effected Predicates)
Earliest attested meanings	causation *per se*	letting and indirect causation in general
Register variation	more tolerated in formal registers	more favoured in informal registers
Geographic variation	more tolerated in Belgian Dutch	more favoured in Netherlandic Dutch

Table 2.1. An overview of the semantic, formal and extralinguistic features of *doen* and *laten,* according to the previous studies.

Chapter 3. Data and method

This chapter describes new tools for a corpus-based multivariate analysis of language semantics, which can be used to explore semantic dimensions, areas, and constellations of usage patterns of one or more words or constructions in any number of language varieties. First, I introduce the theoretical and methodological arguments that support this method. Next, I present the large-scale lectally diverse corpus that was used in this study. A section is devoted to the variables that the data were coded for. Finally, I describe the algorithm of the statistical analysis step by step.

3.1. A theoretical and methodological rationale of the approach

The approach applied in this study is radically bottom-up and usage-based. It is mainly inductive, as most other empirical corpus-based studies in Cognitive Semantics. The main reason for that is the multidimensional character of meaning, which is difficult to capture with a limited set of hypotheses (cf. Gilquin 2010: 8–9). As Wierzbicka (1988: 240) notes, the semantics of specific causative constructions across different languages is more fine-grained and language-specific than can be captured by binary oppositions, such as "direct/indirect causation", "contactive/distant causation", and "strongly coercive/weakly coercive causatives". This means that there may be additional differences between *doen* and *laten*, which stem from peculiar entrenched situations and usage schemata and should be taken into account. In this study I describe the semantic structure of *doen* and *laten* on the basis of a large set of linguistic features, unrelated to any hypothesis, and then try to find the important dimensions in this semantic space. Whether these dimensions reflect directness or indirectness or any other distinction mentioned previously, is then an empirical question. I also explore more fine-grained distinctions between the constructions in local conceptual regions.

Another important issue concerns the nature of semantic evidence. Bottom-up distributional approaches, like the Behavioural Profiles or Semantic Vector Spaces (see Chapter 1, Section 1.3.2), are based on the assumption that the speaker's knowledge of language involves knowledge of the distributional contexts where words and constructions are used. Some linguists and philosophers, including Wittgenstein, go as far as to suggest that meaning IS context of use (see Stefanowitsch 2010: 368–370). A less radical interpretation of the role of contextual clues is that the latter merely reflect the conceptual structures associated with linguistic forms. It has been shown that the situational clues given in the context of a new word or construction are crucial in establishing their meaning, especially when learning verbs and non-basic vocabulary, which normally happens without immediate access to referents (Dąbrowska 2009: 201–202). Some distributional models applied in psycholinguistic research, such as Latent Semantic Analysis (Landauer and Dumais 1997) and Hyperspace Analogue to Language (Lund and Burgess 1996), have been quite successful in performing human lexical tasks, for example, TOEFL tests on synonyms. Latent Semantic Analysis, for instance, arrives at 300 dimensions that are sufficient to distinguish between all words in a vocabulary of an average high school graduate (Landauer and Dumais 1997). However, it still remains to be demonstrated that these dimensions represent some underlying cognitive reality (Murphy 2002: 430).

In the present study, similarity between exemplars is measured with the help of a large set of semantic and other features. The feature-based approach to similarity is the most influential one in the psychology of categorization. For example, Estes' (1994) Array Model framework is based on the assumption that objects and events are stored in memory as vectors containing a list of attributes, e.g. values on perceptual dimensions or 'on/off' categorical features, although it is not yet clear how these features are 'coded' by people in the case of abstract linguistic categories. In the present study I assume that the contextual features of constructional exemplars reflect, albeit indirectly, the properties of the experience verbalized with the help of the constructions and that these properties can serve as the ground for establishing similarity between the constructional exemplars. The semantically relevant distributional information can include both abstract semantic features and formal collocations (cf. Gries 2006; Dąbrowska 2009). Bybee writes,

> “Exemplar representations are rich memory representations; they include, at least potentially, all the information a language user can perceive in a linguistic experience. This information consists of phonetic detail, including redundant and variable features, the lexical items and constructions used, the meaning, inferences made from this meaning and from the context, and properties of the social, physical and linguistic context.” (Bybee 2010: 14).

Because it is impossible to know a priori which contextual features are semantically relevant, I included in the analysis all the features that I was able to code with a sufficient degree of objectivity and precision. The resulting 35 semantic, morphological, lexical and other variables are introduced in Section 3.3.1 of this chapter and described in Appendix 1. Note that the variables are categorical, i.e. they contain different values, or levels. In contrast, psychological models of categorization normally contain binary features, such “has feathers” or “is shiny”.[1]

[1] Although it is possible to binarize the categorical variables used in this study, this would lead to a

Finally, it is important to mention the differences between the present approach and the other quantitative distributional methods. While my method is in some respects similar to the Behavioural Profiles approach by Divjak and Gries (see Chapter 1, Section 1.3.2), the starting point of my data exploration phase is the matrix of dissimilarities between the exemplars, not pre-defined senses or lexical items. This allows me to model polysemy in an entirely bottom-up fashion, whereas Gries (2006) clusters the senses which are defined a priori on the basis of a lexicographer's intuition. The method applied here thus charts the path to fully bottom-up lexicography. As far as near-synonymy is concerned, my approach helps to identify local relationships between near-synonyms, which are especially important for polysemous items. They prevent a researcher from overgeneralizing over the entire sets and show which senses are distinctive of the categories, and which are shared by them.

On the other hand, the present approach is similar to token-based Vector Space Models, which model similarity between exemplars (tokens) in a hyperdimensional space (Schütze 1998). To my knowledge, this approach has been applied only to words, and the similarity of the exemplars is normally based on the distributional properties of the other lexemes that occur in the exemplars' immediate context. In this study, I use a smaller set of manually coded semantic, syntactic and other variables, which describe the context directly. These features also enable me to interpret the exemplar space in semantic terms more directly than it can be done in Semantic Vector Space models.

very high impact of very few variables with many specific levels. This is why this option was not chosen.

3.2. Corpus design

For all analyses presented in this study I use a large data set created from corpora of two varieties of Dutch: the one spoken in the Netherlands and the one used in the Dutch-speaking part of Belgium (Flanders). I used the following three registers, which display conspicuous functional and linguistic differences (e.g. Biber and Conrad 2009):

- spontaneous face-to-face conversations, which constitute a part of the *Corpus Gesproken Nederlands* (*CGN*) – a corpus of spoken Dutch (Oostdijk 2002). The conversations were conducted and recorded in the homes of the volunteers in the period from 1999 to 2003. They represent the most informal type of communication in my data. Such conversations are arguably the most fundamental register of communication (e.g. Biber and Conrad 2009), but they are frequently underrepresented in the current research (Newman 2010: 84);
- postings on the Usenet – online thematic discussion groups, which are currently located at groups.google.com. This subcorpus represents rather informal computer-mediated written communication in a group of registered users on a specific topic over a long period of time. The Usenet was a precursor of contemporary Internet forums, and has seen a decrease in popularity in recent years. The subcorpus represents groups discussing politics (nl.politiek; be.politics), economy (nl.beurs; nl.financieel.beurs, nl.financieel.bankieren; be.finance), football (nl.sport.voetbal; be.sport.football) and music (nl.muziek; be.music). The number of tokens per topic and country was approximately the same. The data that I used cover the years 1997–2010. This subcorpus was neither tagged nor lemmatized, but the messages were stripped from meta-

information, such as the date and time of publication and user's details. Some postings from the most prolific users were discarded to maintain balance between the speakers. Because of massive amounts of spam, all posts with more than one percent tokens that belonged to the English and French function-word stoplists were disregarded. A manual check proved that to be a reasonable threshold;

- newspaper articles from Twente News Corpus (Ordelman et al. 2007) and Leuven News Corpus,[1] two large corpora of Dutch and Belgian newspapers, syntactically parsed with the help of the Alpino parser (Bouma et al. 2001). I used a sample from the years 2001 and 2002. For this period, the articles in the corpora are provided with keywords. These keywords allowed me to identify the topic of the articles and select equal samples of texts about politics, economy, football and music, to match the data from the Usenet.

The structure of the corpus is shown in Table 3.1. The numbers correspond to the number of tokens.

	Spontaneous face-to-face conversations	**Usenet**	**Newspapers**
The Netherlands	1 747 789	1 330 880	1 308 447
Belgium	878 383	1 334 593	1 337 785

Table 3.1. Number of tokens in the subcorpora used in the study.

The next step was to find all observations with the causative auxiliaries *doen* and *laten*. Observations from the spoken data were retrieved with the help of a Python script with regular expressions. I searched for all occurrences of *doen* and *laten* (in all word forms) followed

[1] Leuven News Corpus is a large (1.1 bln. words) corpus of Belgian (Flemish) newspapers created by the Quantitative Lexicology and Variational Linguistics research unit at Katholieke Unversiteit Leuven.

by an infinitive at any distance in the same sentence, using the available part-of-speech tags and lemmata. No such information was available for the Usenet postings, where I had to use regular expressions to search for all forms of *doen* and *laten* followed by a potential infinitive at any distance on the right. The newspaper data were processed with the help of an XML parser.[1] The results showed that the syntactic parsing information was quite reliable in terms of precision, i.e. errors were infrequent. The recall (i.e. how many occurrences were found from the total population of the causative *doen* and *laten*) was difficult to evaluate because the causatives (especially *doen*) are too infrequent in the corpora to allow for a reliable manual check.

The automatically extracted observations were then checked manually to avoid spurious hits. Historically related but different structures with the optative or adhortative *laten*, e.g. *laten wij eerlijk zijn* "let us be honest", were excluded, and so were the sentences with the periphrastic non-causative *doen*, which were especially frequent in the spoken Netherlandic data. Every occurrence of a causative auxiliary was considered a single instance, regardless of how many Causers, Causees or Effected Predicates the instance contained.

The total number of observations with *doen* or *laten* was 5762. Most of them (5031, or 87.3%) contained *laten*, and only 731 (12.7%) were with *doen*. Table 3.2 shows the frequencies of exemplars of two constructions found in the lectal samples after the cleaning-up.

	Spontaneous face-to-face conversations	**Usenet**	**Newspapers**
The Netherlands	811	1154	1201
Belgium	430	1156	1010

Table 3.2. Number of exemplars of the causative *doen* and *laten* in the corpus.

[1] The parser was written by Dirk Speelman and Kris Heylen, to whom I owe my thanks.

I also found three instances of *doen* and *laten* used together in the same context in the spoken data and the Usenet. One of them is below.

(1) *maar ha*a Rob mag ook helemaal niet zo heel veel*
but ha*a Rob may also completely not so very much
hè als dat Andrélaat doet geloven.
hè as that Andrélets does believe
"But Rob may also not so very much, hey, as André lets makes believe." (*CGN*, fn007963)

These three cases, which were also opaque in other respects, were discarded, although they are quite informative as symptoms of the speaker's uncertainty (cf. the discussion of *doen/laten geloven* in Section 2.2.4). The rest of the observations were then coded manually for a number of variables that are described in the following section.

3.3. Variables

Because the aim of this study is to describe the semantics of *doen* and *laten*, the main variable, of course, represents the speaker's choice between the two auxiliaries in every context. It is used as the response variable in many analyses presented here. Due to this methodological choice, the variable did not take part in shaping the exemplar space, although one might argue that the phonological form of the category itself is an integral part of the concept (cf. Abbot-Smith and Tomasello 2006). The rest of the variables constitute the contexts where *doen* and *laten* are used. They are described below.

3.3.1. Local contextual variables

As shown by Gries (2006), Dąbrowska (2009) and Divjak (2010b), all kinds of contextual clues – abstract semantic, abstract syntactic and lexical collocational information – can be important in a semantic representation. In Semantic Vector Space models, it is frequently a combination of lexical and syntactic information that gives the best results. This is why I use a mixture of different indicators in this study. These variables are subdivided into several types:

- features of the Effected Predicate (different types of valency, formal constituents, source and target semantic domains of the caused event denoted by the Effected Predicate). The specific lexemes were also taken into account because of the collocational effects described in Section 2.2.4;
- features of the main nominal participants – the Causer, the Causee and the Affectee (if available): syntactic expression, part of speech, definiteness, semantic class, grammatical person and number. I also coded whether the Causee was performing the caused event intentionally, and whether it was undergoing change or causing a change of another entity;
- other features related to the construction (syntactic function, presence of modal verbs, negation, adverbial modifiers);
- features related to the entire clause (clause type, grammatical mood and tense);
- features related to the entire sentence as a communicative unit (sentence type).

A full list of the 35 variables and their values is provided in Appendix 1. The choice of variables was determined only by practical methodological reasons (although some variables have proven to be useful in the previous studies). I chose the variables which were either formally identifiable, or obvious, or could be coded with the help of a simple test. Those for which it was problematic were discarded, for example, the aspectual class of the caused event. Missing values were allowed. Of course, some of the variables may appear to be very strongly correlated, but the contextual clues that a language learner is exposed to are highly redundant, too.

3.3.2. Global contextual variables

As the previous studies demonstrate, the national variety of Dutch and register have important effects on the probability of *doen* or *laten* in a context. In my study I explored the effects of the following lectal (global contextual) variables: Country (Belgium or the Netherlands) and Register (spontaneous face-to-face conversations, online postings or newspaper articles). These features were not used in determining the similarity between the exemplars in this study, although it would perfectly reasonable to use them, especially the register, as general information about the communicative situation that might be relevant for the speaker. The reason for not doing so was purely methodological. I wanted to explore the three-way relationships between the auxiliary, the semantic context and the different types of lects, in order to disentangle the semantic and lectal sources of variation.

3.4. Quantitative analysis procedure

In this subsection I describe the steps of the quantitative analysis applied in this study, according to the general outline presented in Chapter 1, Section 1.3.2.

Stage 1: collecting the data. The process of the data collection and coding was described above. The result of this stage is a matrix with observations (exemplars) as rows, and variables as columns. Table 3.3 shows the structure of an imaginary sample with imaginary variables.

	Var 1	Var 2	Var 3	...	Var *j*
Exemplar 1	A	“yes”	X	...	...
Exemplar 2	A	“yes”	Y	...	...
Exemplar 3	B	“no”	X	...	...
...	...	...	...	...	...
Exemplar *i*	...	...	...	...	...

Table 3.3. The structure of data that serves as the input for the analysis.

Stage 2: exploring the data. The initial data matrix of local contextual features is used then to obtain a matrix of distances between the exemplars. The distance matrix will play a fundamental role in the subsequent exploratory analyses. The distances represent dissimilarities between the observations, which are based on the dissimilarities between their linguistic contexts. The distances/dissimilarities are calculated with the help of Gower's general coefficient of similarity for all kinds of variables, including categorical ones (Gower 1971). It is implemented in the `daisy` function in the `cluster` package (Maechler et al. 2005) in R, an environment for statistical analysis and a programming language (R Development Core Team 2011).

Gower's coefficient is a straightforward measure of (dis)similarity. In its most general form, it looks as follows (Gower 1971: 861):

$$S_{ij} = \sum_{k=1}^{v} s_{ijk} w_k \Big/ \sum_{k=1}^{v} \delta_{ijk} w_k$$

where S_{ij} is the general similarity between exemplars *i* and *j*, s_{ijk} is the similarity between *i* and *j* with regard to the variable *k*, and δ_{ijk} is the possibility of making comparison between *i* and *j* regarding *k*. δ_{ijk} equals 1 if there are no missing values, and 0 otherwise. Finally, w_k is the weight of the variable *k*, which can be specified by the researcher. By default, all weights are equal to 1. The distances between exemplars are calculated by subtracting the similarity score from 1.

The result of applying this measure to the imaginary data set is shown in Table 3.4. The matrix should be read like a table of distances between cities in a geographic atlas. The zeros on the diagonal mean that there is no distance between an exemplar and itself. 1 means the maximal possible dissimilarity between the objects. It corresponds to the similarity score of 0. Exemplars 2 and 3 have distance 1 because all values that these exemplars have are different (see Table 3.3).

	Exemplar 1	Exemplar 2	Exemplar 3	...	Exemplar *i*
Exemplar 1	0	0.33	0.66	...	...
Exemplar 2	0.33	0	1	…	...
Exemplar 3	0.66	1	0	...	...
...	…	…	…	…	...
Exemplar *i*	…	...	...	...	0

Table 3.4. The distance matrix based on Gower's similarity coefficient for the data in Table 3.3.

To explore the relationships between the exemplars, several sets of analytical procedures can be applied. First, one can study the between-exemplar distances on their own. Second, the exemplar space can be represented as a cloud in a low-dimensional space in order to establish the semantic areas and dimensions and evaluate the density of exemplars in different regions of the space. The third approach shows the hierarchical structure of the observations in a tree-like representation. These complementary approaches are discussed below.

A. Distances between exemplars. The distances between exemplars can be used to evaluate the intracategorial prototypicality of exemplars, similar to their family resemblance. The ones with the smallest average distance to all other members can be considered the most prototypical, whereas those with the largest distance are the least representative. In our artificial example, Exemplar 1 would be the most central, and Exemplar 3 would be the least typical of the category. One can also measure the intercategorial cue validity of an exemplar by dividing its average distance to the members of its own category by the average distance to all exemplars in two categories. The lower the score, the higher the cue validity.

At the level of categories, the average distance between all exemplars of one category can serve as an operationalization of the semantic variability of the category, its abstractness, and its size, depending on the perspective. One can expect these properties to be related to productivity, although they are not identical to it (cf. Bybee 2010: 91–94).

B. Semantic maps. One can represent the exemplars in a low-dimensional space with the help of Multidimensional Scaling (MDS) applied to the distance matrix. Throughout the study I use the iterative majorization algorithm implemented in the `smacof` package in R (de Leeuw and Mair 2009), which is a relatively simple and powerful technique, which usually provides good results (Borg and Groenen 1997). The two-dimensional map for the imaginary data set is shown in Figure 3.1. One can see that the maximal distance is between Exemplars 2 and 3, in accordance with what one can see in the distance matrix, whereas Exemplar 1 and Exemplar 2 are maximally close. The map also shows that

Exemplar 1 is the most central, and Exemplar 3 is the most peripheral, as was established previously. The family resemblance is related to the position of the exemplars in the centre or periphery of an MDS map because the exemplars that are the closest to all others are normally positioned in the centre of the solution. On the other hand, the cue validity (intercategorial salience) of an exemplar can be estimated with the help of the map, too, as the distance from the exemplar to the members of the contrasting category.

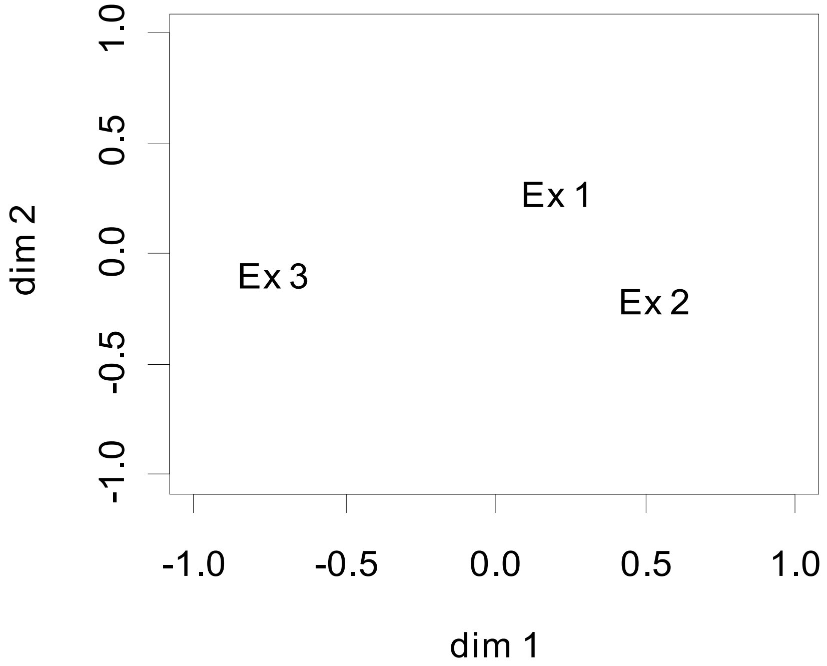


Figure 3.1. A MDS solution for the data in Table 3.3.

Once we have established the structure of the data, the next important step is to interpret the groups of exemplars. One can do it manually, of course. For instance, Exemplars 1 and 2 in the data set have the same values of Variable 1 ("A") and Variable 2 ("yes"), whereas Exemplar 3 has "B" and "no", respectively. Variable 3 is less distinctive of Exemplars 1 and 2 as a cluster because Exemplar 3 shares the value "X" with Exemplar 1. However, applying this method for hundreds of observations would be difficult. Two more technical solutions are available:

a) plotting the levels of variables according to the positions of the exemplars with these levels. This allows one to find the important dimensions and clusters of exemplars that share the same features;

b) a more formal procedure – a series of ANOVAs or linear regressions on the coordinates of the MDS solution with the variables as predictors (factors). One can compare the *F*-scores or measure of explanatory power and see which variables are the most strongly associated with the MDS dimensions.

C. Hierarchical structure of usage patterns. Although the MDS map clearly shows two groups (Exemplars 1 and 2 vs. Exemplar 3), it may be desirable to perform cluster analysis automatically when there are many exemplars. A large number of various clustering algorithms exists (Everitt et al. 2001). In this study, I use hierarchical agglomerative clustering (Ward's minimum variance method) implemented as the `hclust` function in R.

Hierarchical clustering is arguably the most popular clustering technique. In the beginning of the algorithm, every object (in this case, every exemplar) in the distance matrix forms its own cluster. Next, the algorithm joins the two objects with the minimal distance between them, and then proceeds iteratively, joining objects and small clusters until they all form a single megacluster (the 'root'). The analysis can be represented as a dendrogram with 'branches' and 'leaves' (see Figure 3.2).

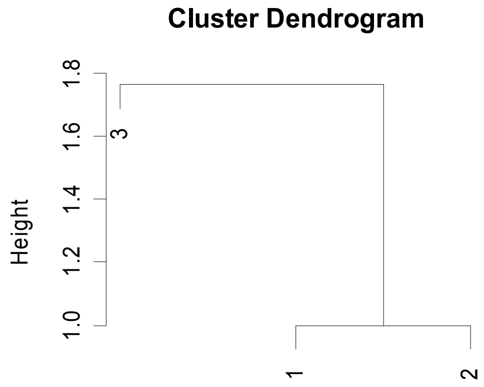


Figure 3.2. A dendrogram of the data in Table 3.3 (hierarchical cluster analysis).

The height at which the objects merge reflects the distance between them, and thus indicates their dissimilarity. The higher they merge, the shorter the 'stem' they hang from. Therefore, short stems indicate a merger of relatively heterogeneous exemplars, whereas long stems suggest distinct homogeneous clusters. Applying the method to the data in Table 3.3, one obtains a dendrogram like the one shown in Figure 3.2. As one could expect, Exemplars 1 and 2 form one cluster, and Exemplar 3 is set apart.

When one has to explore a large dendrogram with several hundreds of observations, it is difficult to get a grasp of the structure by inspecting the observations manually. Instead, one can find distinctive features of a pair of clusters formed by each split with the help of methods similar to the ones used in the Behavioural Profiles approach. There are several possible ways of doing this. One way is to calculate the average profiles of each cluster or category as a vector with the relative frequency of each semantic feature in the cluster, and next to study the largest differences between the relative frequencies between the two clusters (Gries and Otani 2010). A more sophisticated approach is used by Divjak (2010b), who takes into account the distribution of every variable across the clusters, in a way that resembles the *t*-test (albeit without hypothesis-testing power). This approach has its advantages when one compares more than two clusters. Since I deal with binary splits, the simpler solution is used. The differences in proportions can be interpreted as a simple operationalization of the cue validity of the features, and the proportions themselves serve as indicators of the features' weight in the category structure.

Semantic phenomenon	Operationalization
Semantic similarity between exemplars	distance between exemplars in the distance matrix (inversely correlated); closeness on the MDS map
Semantic similarity between categories	overlap of the corresponding areas on the MDS map
Semantic dimensions	dimensions identified in the MDS solutions, including the principal dimensions (axes), diagonal and non-linear patterns
Semantic autonomy	distance of a cluster or an exemplar from the rest of the exemplars of the same category
Semantic variability of the category, or its size	average distance between all exemplars of the category
Entrenchment of a sense or a subschema	density of the corresponding group of exemplars in the MDS map
Intracategorial family resemblance of individual exemplars	inversely correlated with the average distance between the exemplar and the other members of the same category; located in the central part of the MDS representation of the category
Intercategorial cue validity of individual exemplars	inversely correlated with the average distance between the exemplar and the other members of the same category divided by the average distance to all exemplars of several different categories; located in the part of the MDS map maximally distant from exemplars of the contrasting category (-ies)
Intracategorial weight of individual features	proportion of the feature among the exemplars of the category
Intercategorial distinctiveness of individual features	difference in proportions of the feature among exemplars of contrasting categories

Table 3.5. Operationalization of different semantic phenomena in the present study.

Whereas most other empirical studies have focused on the cue validity of specific features, the above-mentioned techniques also provide information about the objects (exemplars, senses and categories) and their semasiological and onomasiological salience. Table 3.5 summarizes the operationalizations of these and some other semantic phenomena in the present study. Note that the word 'semantic' is used in a broad sense, and includes pragmatic and other phenomena of language use.

Stage 3: confirmatory tests. The distinctiveness of the inter-cluster and inter-categorial features can be confirmed with the help of a logistic bi- or multinomial regression (see Chapter 1, Section 1.3.2), which allows one to establish how well these features discriminate between the clusters or categories. This is a popular way of testing the validity of exploratory analyses. In addition, one can also test the lectal differences found at the previous stage. In cases of data sparseness and complex interactions it is preferable to use non-parametric techniques, such as conditional inference trees and random forests (Tagliamonte and Baayen, Submitted). These methods will be introduced in Chapter 6.

3.5. Summary

This chapter has outlined the data and quantitative techniques that are used in the rest of the study. The analyses are based on a large lectally diverse corpus of present-day Dutch, coded for several dozens of semantic, syntactic and other features. The central part in the analyses is played by the similarities between the exemplars, which are operationalized as the similarities between their contexts. These similarities can be transformed into spatial distances and represented visually with the help of Multidimensional Scaling maps, which allow one to explore the semantic structure of the categories. To arrive at the hierarchical network of the constructions and a finer semantic representation, I use cluster analysis. A range of simple and advanced techniques is available to establish the prototypical and distinctive features of categories and usage patterns. The next chapter, which presents a semasiological analysis of the causative construction with *doen*, shows this approach in action.

Chapter 4. Semantic and lectal variation of the causative construction with *doen*

In this chapter I apply the approach outlined in the previous chapter to model the semantics of the Dutch causative construction with *doen*. More specifically, I explore the family resemblance relationships between the exemplars (Section 4.1), dimensions that are relevant for the semantic structure of the construction (Section 4.2) and perform a cluster analysis of the exemplars to describe the main senses of the construction and its most important low-level schemata (Section 4.3). The chapter also presents an analysis of geographic and situational variation in the use of *doen* (Section 4.4). I show that the variation is not only quantitative, but also qualitative, and that it involves differences in the entrenchment of specific senses and collocations in different varieties of Dutch. Section 4.5 offers a brief summary of the main results.

4.1. Intracategorial salience of exemplars of *doen*

At the first step of the analysis, the data set with 731 exemplars of *doen* coded for the above-mentioned 35 local contextual variables was transformed into a matrix of distances between the exemplars with the help of Gower's similarity measure. The range of distances between the pairs of different exemplars was from 0 to 0.79: this means that there were pairs that shared all features, but no two exemplars had different values for every variable. The average distance between the exemplars was 0.36 with the standard deviation of 0.11, which suggests a moderately homogeneous category. I also calculated the average distance from every exemplar to the other exemplars in order to see which ones would display the highest family resemblance with the other exemplars by sharing the largest number of features with them. Taking the top fifty most prototypical exemplars that were located at the minimum distance from the other exemplars, I found that all those exemplars had abstract Causers, intransitive Effected Predicates, and Causees that did not act intentionally. Most exemplars from this top fifty referred to mental events (source and target domains),[1] had human individual Causees, contained the Effected Predicate *denken* "think" and preposition *aan.* The prominence of the lexemes can be explained by the high frequency of the collocation *doen denken aan* "remind of". In addition, the overwhelming majority occurred in declarative sentences and main clauses with indicative mood and present tense, were used in the predicative function, did not have any adverbial modifiers, negation or modal verbs, and had an explicit definite 3rd person singular Causer, which was not coreferential with the other participants and did not possess them. All these features were the most frequent values of the corresponding variables in the entire data set, except for the prepositional complement with *aan*. Consider (1), an exemplar that had the highest similarity score with all other observations.

(1) *Hee, dat doet me denken.*
Hee that makes me think
"Hey, that makes me think." (*nl.sport.voetbal*)

This suggests that affective causation might be a good candidate for the most prototypical sense of *doen*. However, this effect is due to the high frequency of the collocation *doen denken aan*.

Although the exemplars that were the most distant from the others

[1] The source semantic domain reflects the domain of the caused event as specified by the Effected Predicate in the literal sense, whereas the target domain describes the figurative meaning (the actual referent). These domains may not coincide in the case of figurative expressions. The figurativeness of caused events was determined conservatively: non-creative uses attested in dictionaries were not considered figurative, unless they were marked as such. See examples in Appendix 1.

display a greater diversity of features, most of them have human Causers and express social caused events. The one with the lowest similarity score also contains reflexive and possessive pronouns and a transitive Effected Predicate:

(2) (…) *tussen* *je kinderen* *je* *doen* *verzorgen* *en*
between your children you do look-after and
(…) *regelmatig* *bezoek* *hebben* *daarvan*.
regular visit have of-them
"(...)between having your children look after yourself and being visited by them." (*CGN*, fv400655)

In (2) and other untypical exemplars with low similarity scores, it would be perfectly natural to use *laten*. This use of inducive interpersonal *doen* seems to be archaic and/or dialectal (see Section 2.3 of Chapter 2).[1] In this case, low family resemblance to the members of its own category means high typicality with regard to the contrasting category (cf. Rosch and Mervis 1975). More on this will follow in Chapters 6 and 7.

4.2. The exemplar space of *doen*

Next, the distances between the exemplars were used for a Multidimensional Scaling analysis, which allowed me to represent the exemplar space of *doen* in a two-dimensional solution. The stress value of the solution was approximately 0.07. This means that the map represents the (dis)similarities between the exemplars in a reliable way. The solution is displayed in Figure 4.1. The points are the exemplars of *doen*. The closer they are to one other, the more semantic and other features they share. The

[1] The speaker and her interlocutor were two West-Flemish women over 50.

exemplars in the centre have the above-mentioned features that were typical of most exemplars of *doen*. However, the centre of the plot is sparsely populated. This suggests that the general schematic meaning of the category should be quite schematic because it is not represented by many specific exemplars.

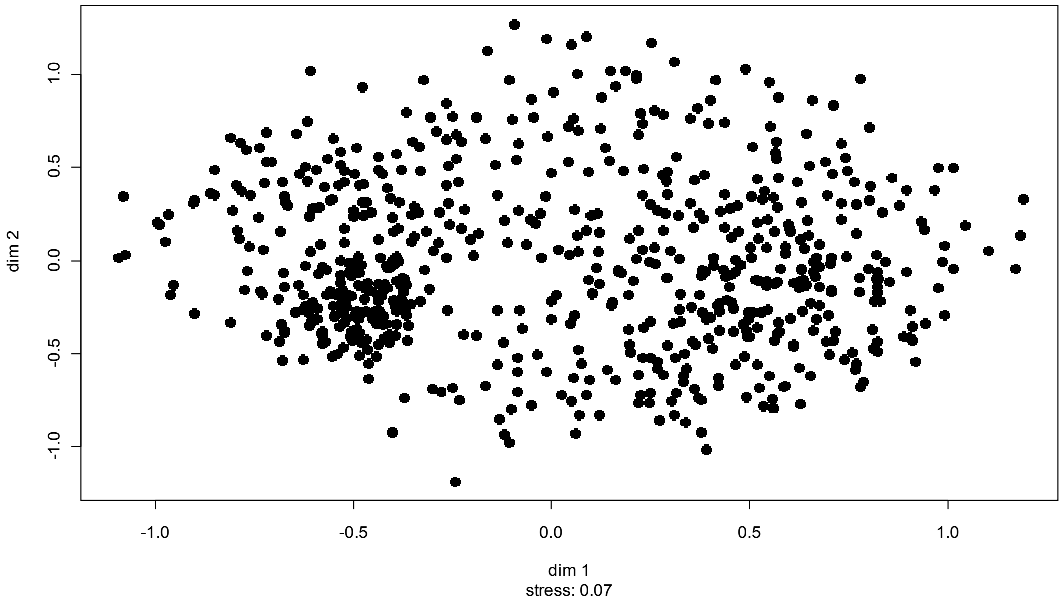


Figure 4.1. MDS representation of the exemplar space of *doen*.

The dimensions of the solution can be interpreted by applying the two methods mentioned previously. The first method is to map the levels of the variables on the MDS solution, as shown in Figure 4.2. This procedure ensures that there are no non-linear semantic distinctions that one could miss by restricting the analysis to the two dimensions (cf. Borg and Groenen 1997). The second, more objective approach, allows the interpretation of the two principal dimensions of the solution with the help of a series of Analyses of Variance (ANOVA). For each categorical variable, I calculated the *F*-score, which indicates that the differences between the average coordinates for different levels of the variable are significantly greater than the differences between the observations with the same value. I also used the R^2 measure, which shows how well the variable can explain variation in the positions of the exemplars in a linear regression model.

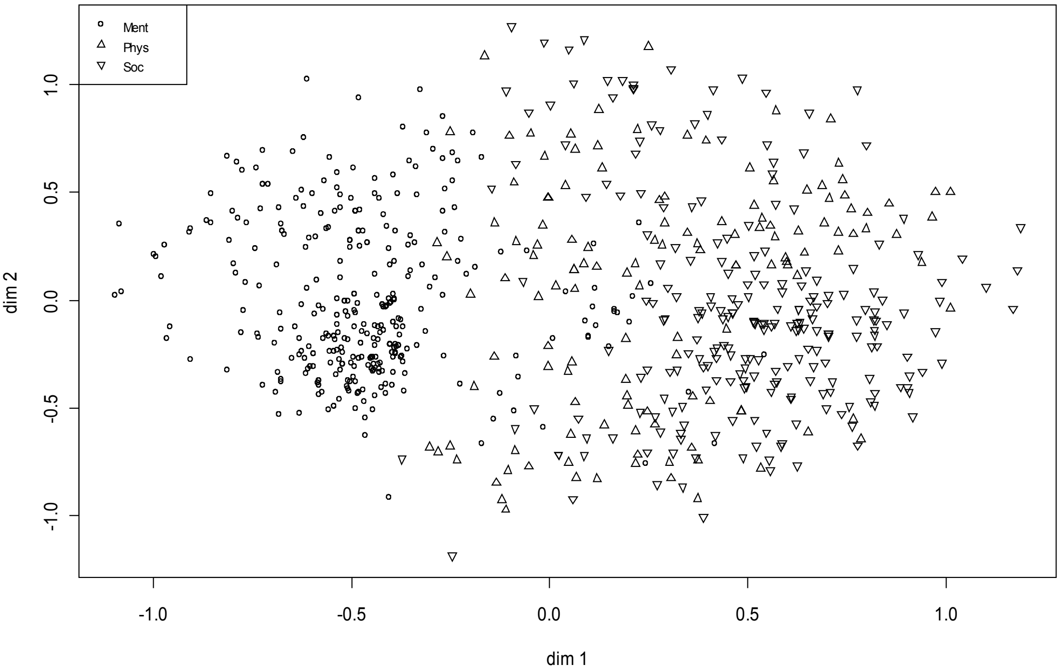


Figure 4.2. Distribution of the source semantic domains of caused events in the exemplar space of *doen*.

Both approaches show that the most important semantic distinctions coincide with the two principal dimensions of the MDS solution. The first (horizontal) dimension corresponds to the distinction between mental (on the left), and physical or social (on the right) caused events, in the literal and figurative interpretations (the *F*-scores were 970 and 841.4, respectively, on 2 and 727 d.f., $p < 0.001$). These variables also explain the variance in the coordinates of the exemplars the best (the R^2 values were 0.73 and 0.70). Some of the other variables strongly associated with this distinction are the semantic class, part of speech and grammatical person of the Causee: mental caused events normally require human Causees, which, unlike the inanimate ones, can be expressed with the 1st and 2nd person pronouns. Compare (3) and (4), which represent two points with very large absolute values on the horizontal dimension:

(3) *Dat wil jij ons hier nu al beriechtenlang*[1]
that want you us here now already messages-long
heel krampachtig doen geloven, maar je krijgt iets
very forcefully do believe but you get something
van een roepende in de woestijn.
of a crying in the desert
"This is what you have been trying very hard to make us believe in many messages, but you are like a voice crying in the wilderness." (*nl.politiek*)

(4) *Maar waarom kunnen zij niet gewoon die*
but why can they not simply those
twee aparte ziekenhuizen verder doen draaien?
two separate hospitals further do run?
"But why can't they simply continue running these two hospitals separately?" (*CGN*, fv400506)

In the first example, the caused event is mental, and the Causee is a group of people. It is an example of a failed direct interference into a human mind (see the discussion in Section 2.2.4). The second sentence is an example of a social caused event.

It is important to highlight that these types of causation are not entirely separate. In the middle of the map there are quite a few intermediary cases. Most of them involve a mental caused event conceptualized as a physical event, e.g. (5):

(5) *Ben altijd nieuwsgierig of mijn info ergens*
am always curious if my information anywhere

[1] The original orthography of the examples is preserved.

een belletje doet rinkelen.
a bell does ring
“I'm always eager to know if my information rings a bell anywhere.” (*nl.muziek*)

The second (vertical) dimension separates the caused events that happen regardless of the Causee's intentions (bottom), from those that involve the Causee's intentional actions (top). The *F*-statistic was 154.6 on 2 and 728 d.f, $p < 0.001$, $R^2 = 0.30$. The other variables associated with this dimension were transitivity (intransitives vs. transitives) of the Effected Predicates, as well as the grammatical number and definiteness of the Causee (singular opposed to plural and definite opposed to indefinite, respectively). These features can be interpreted in terms of the semantic role of the Causee. The non-volitional Causees in intransitive causative constructions are the end point of the energy flow and are therefore very patient-like, whereas the volitional Causees in transitive constructions are less affected and more active. A close inspection of the variable-specific maps shows that the Causees at the top often bring about a change in another entity (the variable *CeRole*), i.e. they have some properties of an agent.[1] In addition, definite and singular Causees are more individuated than indefinite and plural ones. In this way, the former more strongly resemble patients in prototypical transitive constructions than the latter because an action can be easily transferred to them from the agent (Hopper and Thompson 1980: 253). All this suggests that the dimension can be interpreted as the difference between more direct and less direct causation: short causation chains with more patient-like Causees at the bottom and longer chains with less affected Causees at the top. Compare (6), which has an extremely low value on the dimension (it contains a translated

[1] This variable is also associated with the first dimension, with lack of any observable change in the mental part, and a change of the Causee in the non-mental part of the map.

English idiom), and (7), located at the very top of the map (it has already been discussed as the least prototypical exemplar of the construction):

(6) *En een argument dat me (...) over de grond*
and an argument that me over the ground
zou doen rollen van het lachen.
would do roll from the laughter
"And an argument that would make me (…) roll over the floor laughing." (*nl.politiek*)

(7) (…) *tussen je kinderen je doen verzorgen en*
between your children you do look-after and
(…) *regelmatig bezoek hebben daarvan.*
regular visit have of-them
"(...)between having your children look after yourself and being visited by them." (*CGN*, fv400655)

It is worth noting that the dimensions that emerged as a result of the MDS analysis are similar to the most fundamental distinctions in the perception of human actions – namely, publicly observable vs. unobservable actions, and intentional actions vs. unintentional behaviour (Malle 2005). Of course, the semantics of *doen* is broader than causing a human action, but these two dimensions might be fundamental for conceptualization of all events, not only human actions.

In addition to the semantic dimensions, one can also interpret the density of exemplars, as displayed in Figure 4.3. The map shows two large densely populated areas. The most dense cluster is in the mental part of the map. It consists of the exemplars with *doen denken aan* "remind of". Thus, we have evidence of very high entrenchment of this collocational pattern and the corresponding sense. The high density area on the right corresponds to the exemplars with social caused events and mostly abstract Causees, but no dominant lexical pattern. In addition, there is another small region of high density in the mental part above the cluster with *doen denken aan*. It mainly contains observations with *doen vermoeden* "suggest, make one suspect", as in (8):

(8) *Anders dan de forse nederlaag doet vermoeden speelde*
other than the strong defeat does suppose played
Sparta helemaal niet zo beroerd.
Sparta at-all not so terribly
"In contrast with what the heavy defeat suggests, Sparta did not play that terribly." (*AD*, Nov. 2001)

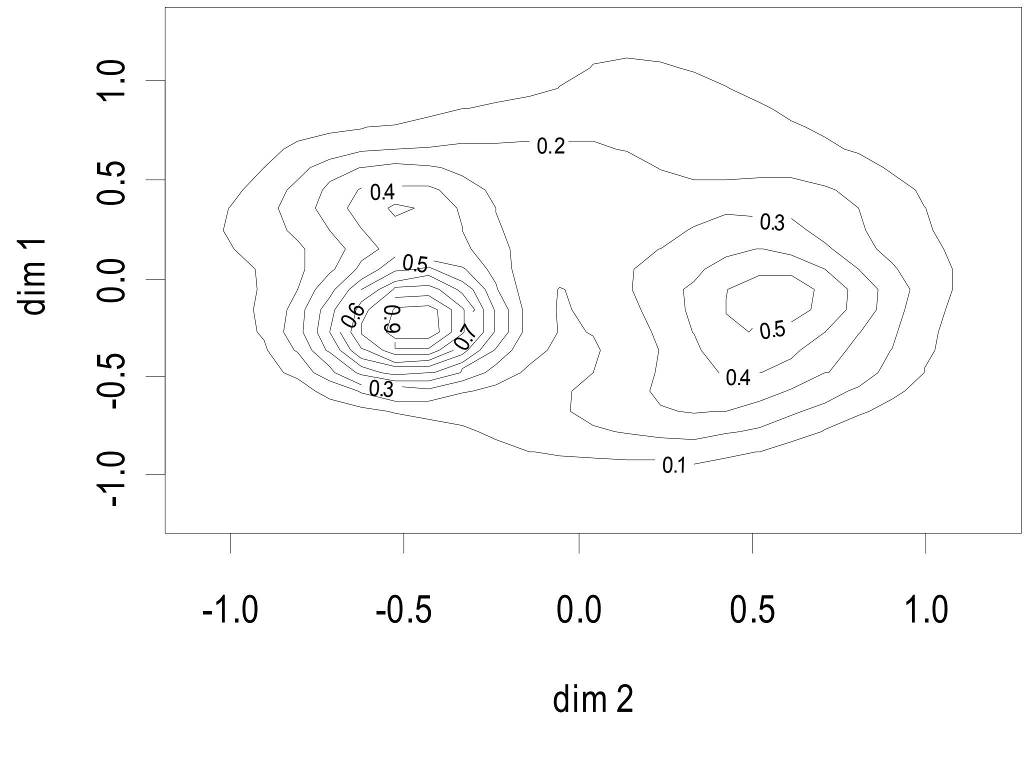


Figure 4.3. Density of the exemplars of *doen* in the different regions of the map.

The areas in the middle, between the main clusters, and in the upper part (indirect causation) have very low density. In semantic terms, density is expected to correlate with entrenchment of the corresponding sense. According to the map, the two most populated large clusters should be more salient than the other regions of the space. Recall that Stukker (2005) proposed two prototypes of *doen*: affective and physical causation (see Chapter 2, Section 2.1).[1] In addition, high density may lead to exemplar effects that should influence categorization. Therefore, we can expect the probability of *laten* in these densely populated areas to be lower than in the other regions.

4.3. Hierarchical cluster analysis of the exemplars of *doen*

In this section I apply hierarchical cluster analysis to model the constructional network of *doen* and describe its main usage patterns. I understand usage patterns as not only traditional senses, which refer to conceptual contents of a category, but also as groups of exemplars with common processing-related, pragmatic and lexical features.

In a hierarchical cluster analysis (see also Chapter 3, Section 3.4) the solution 'grows' from the leaves to the roots, from more similar exemplars to more generalized groups. The leaves themselves are probably less interesting from the cognitive point of view than the more abstract branches because the specific instantiations of constructions normally fade away in the memory with time. This is why I will mainly focus on the large clusters. The top 'splits' that I focus on in fact represent the last amalgamations. In other words, they integrate the most dissimilar groups of exemplars. To interpret them semantically, I used the method applied in Gries and Otani (2010) to behavioural profiles of lexical forms (see Chapter 3, Section 3.4). First, I calculated the relative frequencies (proportions) of the levels of every variable for each cluster in a split. For

[1] Note that 'physical' in Stukker's (2005) study meant involving an inanimate Causer and Causee, thus it also covers most events in the area with social caused events and abstract Causees.

example, in the top split, which is indicated as 1 in Figure 4.4, 97% exemplars of the cluster on the right refer to mental caused events (source domain), as compared to only 5% in the second cluster. Next, the values in the first cluster were subtracted from the corresponding values in the second cluster: 97% – 5% = 92%. I did so for all levels of all variables and then ranked the resulting differences (the absolute values, regardless of the sign). For Split 1, the above-mentioned feature ranks the highest, i.e. it is the most distinctive. It is followed by mental caused events (target domain) and individual human Causees, preferred in the right-hand cluster. Thus, Split 1 corresponds to the semantic distinction between mental and non-mental caused events in the first (horizontal) dimension of the MDS solution.

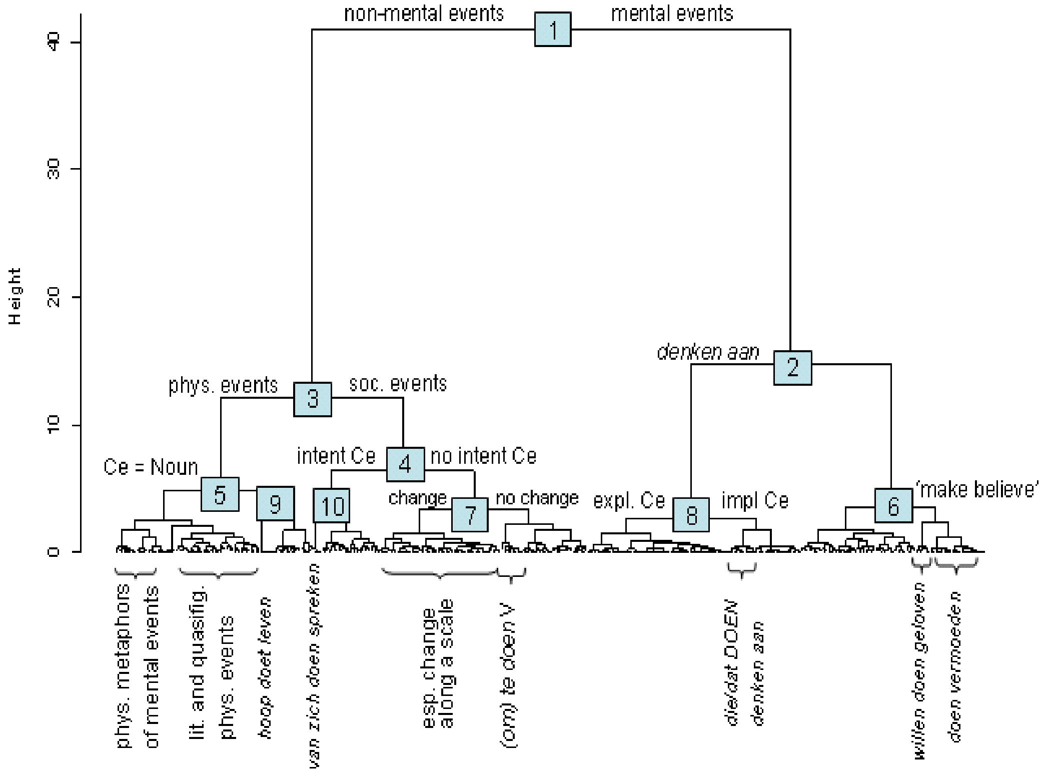


Figure 4.4. Hierarchical cluster analysis of the exemplars of *doen*.

After this, I took the next split and compared the proportions of the variables in the new pair of clusters, and so on. I explored twenty binary splits, and inspected the remaining smaller clusters manually. In what follows I discuss only the 10 top splits and mention the most remarkable smaller subclusters. For brevity, I will discuss only the most distinctive features, unless there are several prominent features that highlight different aspects of semantics.

Split 2 separates the cluster on the left with predominantly *doen denken aan* (97% of all exemplars) from all other collocations. Recall that the relative length of the stem correlates with relative similarity of the observations in this cluster. Following this logic, the exemplars with *doen denken aan* form a quite distinct homogeneous cluster, as could also be seen from the MDS map in the previous section. The cluster then bifurcates in Split 8, forming the clusters with explicit and implicit Causees. The explicit Causee is most commonly the first person singular pronoun *mij* or *me* "me", so the construction is used for sharing subjective associations, as in (9):

(9) *die doet me een beetje aan schilderijen van Jonas*
that does me a bit to paintings of Jonas
denken.
think
"This reminds me a bit of Jonas' paintings." (*CGN*, fn000741)

The other branch of the *doen denken aan*-cluster contains implicit Causees, as in (10). In all of them the focus shifts from personal subjective experience to some objective properties of the Causer, for example the stylistic peculiarities of a piece of music:

(10) *Het recurrente sixties-orgeltje doet denken aan Dylans*
the recurrent sixties-organ does think to Dylan's
Blonde on Blonde.

Blond on Blond
“The recurrent organ from the sixties reminds of Dylan's Blond on Blond.” (*De Morgen*, Oct. 2001)

This group also contains a subcluster with the relative clauses of the type *dat /die DOEN (...) denken aan X* “which reminds of X”. In such cases, the refocusing is evident, since the position in the relative clause makes the construction highly descriptive:

(11) *Evenals het met snauwende rockgitaren omlijste*
just like the with snarling rockguitars framed
"The fear", dat deed denken aan Radiohead ten tijde
The fear which did think to Radiohead at-the time
van The bends.
of The bends
“Just like 'The Fear', framed with snarling rockguitars, which reminded of Radiohead at the times of The Bends.” (*De Standaard,* Feb. 2002)

The right branch of Split 2 (all mental Effected Predicates except for *doen denken aan*) contains Split 6, which separates from all others the situations where it was not clear whether the Causee acted intentionally or not (98% of all exemplars). This coding was applied when the caused event could be seen as both controllable or uncontrollable (cf. D'Andrade 1987). The cluster mostly contained two verbs: *vermoeden* “suppose”and *geloven* “believe”. In its turn, this cluster splits (Split 12, not shown in the dendrogram) into a cluster with the situations when someone tries to make someone believe something intentionally (with the modal verb *willen* “want” in 100% of all cases) and a cluster without such an intention, when certain properties of the Causer trigger some conclusions. Compare (12) and (13):

(12) *Dat wil jij ons hier nu al beriechtenlang*
that want you us here now already messages-long
heel krampachtig doen geloven, maar je krijgt iets
very forcefully do believe, but you get something
van een roepende in de woestijn.
of a crying in the desert
"This is what you have been trying very hard to make us believe in many messages, but you are like a voice crying in the wilderness." (*nl.politiek*)

(13) *Anders dan de forse nederlaag doet vermoeden speelde*
other than the strong defeat does suppose played
Sparta helemaal niet zo beroerd.
Sparta at-all not so terribly
"In contrast with what the heavy defeat suggests, Sparta did not play that terribly." (*AD*, Nov. 2001)

In the cluster exemplified by (13), the meaning is very frequently contrastive. This contrast is expressed by the elements *(niet zo) zoals* "(not so) as", *anders dan* "different from", or *in tegenstelling tot* "in contrast with", which introduce the subordinate clause with *doen*. Both clusters contain exemplars that refer to incorrect information and deceitful appearances, which can mislead the Causee.

The left cluster of Split 6 contains a variety of other mental processes denoted by such Effected Predicates as *besluiten* "conclude, decide", *vergeten* "forget", *dromen* "dream of", *nadenken* "think about", *verlangen* "long for" and *vrezen* "fear". One of the relatively frequent fixed expressions is *het ergste doen vrezen* "fear the worst", as in (14):

(14) *Enkele recente uitlatingen van de chef-logistiek van het*
*s*ome recent comments of the head logistics of the
organiserende comité doen alvast het ergste vrezen (...).
organizing committee do meanwhile the worst fear
"Meanwhile, some recent comments from the head of logistics from the organizing committee make one fear the worst." (*De Morgen,* Jan. 2001)

Let us now go back to the top split and explore the structure of the non-mental cluster. After Split 3, two clusters emerge: one with predominantly physical caused events and the one with social caused events. The 'physical' cluster later divides in Split 5 into two clusters with nominal or pronominal/implicit Causees. The cluster with pronominal or implicit Causees is very heterogeneous and contains several smaller ones, which involve inducive causation with intransitive or transitive predicates, as in (15) and (16), respectively. In both cases *laten* would be more appropriate:

(15) *ze doen u blazen uh en ze pakken*
they do you breathe uh and they take
gewoon uw rijbewijs in.
normally your driving-license in
"They give you a breath test and they normally take in your driving license." (*CGN*, fv400722)

(16) *(...) 'k heb*z ze opnieuw doen vullen want dat*
I have them again do fill because that
waren gevulde tanden.
were filled teeth

"(...) I had them filled again because those were filled teeth." (*CGN*, fv400656)

In addition, this cluster contains a tiny but very homogeneous cluster with the pre-fab sentence *hoop doet leven* "while there is life there is hope (lit. hope makes live)", formed after Split 9.

The left branch of Split 5 contains several interesting clusters: the physical caused events (target domain), as in (17), a cluster with quasi-figurative caused events, which often specify the physical symptoms of a mental event, as in (18), and a few quite distinct clusters with metaphorical expressions denoting mental or social caused events: *een belletje doen rinkelen* "ring the bell", as in (19), *de druppel die de emmer deed overlopen* "the drop that made the cup (lit. bucket) flow over", and *stof doen opwaaien* "raise controversy (lit. raise dust)".

(17) *De bewegende instrumenten worden aangedreven door*
the moving instruments are driven by
een motor, die gelijk ook de kartonnen partituur
a motor which at-the-same-time also the cardboard score
doet draaien.
does rotate
"The moving instruments are driven by a motor, which at the same time turns the cardboard score." (*De Standaard*, Feb. 2002)

(18) *Ook Pepsico kon het beleggershart niet sneller doen*
also Pepsico could the investor-heart not faster do
kloppen.
beat
"Pepsico could not make the investors' hearts beat faster, either." (*De Volkskrant*, Feb. 2002)

(19) *Ben altijd nieuwsgierig of mijn info ergens*
am always curious if my information anywhere
een belletje doet rinkelen.
a bell does ring
"I'm always eager to know if my information rings a bell anywhere." (*nl.muziek*)

The cluster with the social caused events (the right branch of Split 3) is very large. In Split 4, it divides into a cluster with unintentionally acting Causees, and the one with Causees acting intentionally, but doing so under pressure, as in (20):

(20) *Waar haalt enig politiek vertegenwoordiger nog enig recht*
where gets any political representative still any right
om welke belg ook om het even wat op te leggen of
in-order any Belgian no matter what to impose or
dwingend te doen opvolgen?
forcefully to do obey?
"Where does any political representative get any right to impose anything on any Belgian or force him/her to obey?" (*be.politics*)

This cluster also contains a distinct group of exemplars with the frequent expression *van zich doen spreken* "make oneself noticed (lit. make speak about oneself)" (Split 10):

(21) *Batistuta heeft bij AS Roma eindelijk van zich doen*
Batistuta has at AS Roma finally of self did
spreken.
speak

“Batistuta finally managed to get himself noticed at AS Roma.”
(*De Volkskrant*, Oct. 2001)

The cluster with unintentionally acting Causees is further split into the contexts where the Causee undergoes change, and the situations where that does not happen (Split 7). In the cluster without any observable change, one can find a lot of expressions that refer to the perception of a social situation: to count, be noticed, as in (22), or to create a (false) impression, as in (23). These meanings are represented by the verbs *gelden* “count, weigh” and *voorkomen* “appear”:

(22) *Bij Manchester United hebben de aankopen van Sir*
by Manchester United have the purchases of Sir
Alex Ferguson zich eindelijk echt doen gelden.
Alex Ferguson self finally really do count.
“At Manchester United, the acquisitions made by Sir Alex Ferguson have finally really started to take effect.” (*De Morgen*, Oct. 2001)

(23) *Ook doen sites het voorkomen dat concerten al*
also do sites it appear that concerts already
uitverkocht zijn terwijl dat niet zo is.
sold-out are while that not so is
“Also sites make it appear that concerts are already sold out, while this is not so.” (*nl.muziek*)

In the cluster where the Causee undergoes a change, the most frequent type involves Effected Predicates that specify some quantitative change along a scale, as in (24), which, as many similar examples, comes from a newspaper article on an economy-related topic:

(24) *Dat zou de bedrijfswinst van Repsol dit jaar doen*
that would the profit of Repsol this year do
zakken van de verwachte 1,62 tot 1,58 euro per
fall from the expected 1,62 to 1,58 euro per
aandeel.
share
"That would make the profit of Repsol go down from the expected 1.62 to 1.58 euros per share this year." (*De Morgen*, Jan. 2002)

Apart from the quantitative change, one can also find a few examples of influencing the outcome of an event:

(25) *Hij wees er ook op dat het tijd kost om*
he pointed there also on that it time cost in-order
de missie te doen slagen.
the mission to do succeed
"He also pointed out that it cost time to bring the mission to success." (*AD*, Oct. 2001)

To summarize, the analysis of the clustering solution has demonstrated that the semantic domain of the caused event is the most important distinction for the exemplars of *doen*. Therefore, the classification of direct causation into the so called physical and affective causation is supported by the data. However, the transitivity of Effected Predicates, which was mentioned in the typology of causative constructions by Kemmer and Verhagen (1994) is not particularly important for *doen*. Implicitness of the Causee does play a role, but only in the cluster with *doen denken aan*, where it serves for refocusing on objective features of the Causer, and, marginally, in the cluster with physical caused events, where it performs various functions in a range of small clusters. Lexical similarity of the Effected Predicates and even some other slot fillers is extremely important for aggregation both at the low and top clustering levels. The pattern *doen denken aan* forms a huge separate cluster, which amalgamates with its sibling at a significant height.

The analysis shows also shows that different oppositions may be important for different areas of the semantic space. For instance, for mental caused events the role of the Causee in the event specified by the Effected Predicate is irrelevant because they most frequently denote states, which do not involve any change. The distinction between the figurative and literal meanings is more important for physical caused events, because this domain is a frequent source of metaphors for all others.

It is worth mentioning that some of the senses that are encyclopaedically and conceptually similar occur in different parts of the dendrogram. The most vivid example is *iemand iets doen geloven/vermoeden* “make someone believe something” in the cluster with the mental Effected Predicates and *iets doen voorkomen* “make something appear” in the 'social' cluster. Two other frequent expressions found in different branches refer to memory and associations: *aan X doen denken* “remind of X” and *een belletje doen rinkelen* “ring a bell”. One needs a very subtle coding schema based on a fine-grained ontology of events in order to be able to incorporate these similarities in a model. This challenge is left for future research.

Another reservation is that the splits/amalgamations are binary, which does not necessarily reflect the speaker's way of generalizing over exemplars and storing the network in the memory. Another concern is that the type of clustering presented here is 'crisp', that is, it does not allow for multiple probabilistic membership of an exemplar in several clusters. In this regard, it might be useful to combine the clusters with the MDS map, which emphasizes the continuity of the space.

Figure 4.5 shows the five coarse-grained clusters produced after 4 top splits on the MDS solution presented in Section 4.2, to give an idea of the distribution of the most important causation types. The numbers of the clusters stand for the following usage patterns:

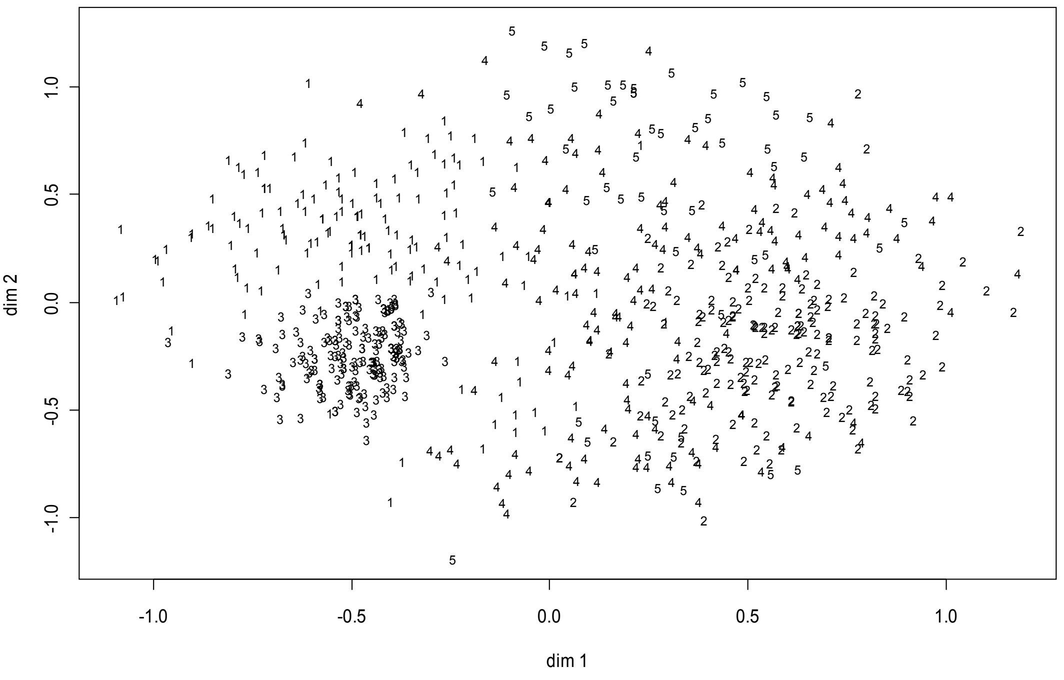


Figure 4.5. Five usage clusters projected on the MDS solution.

1: the big mental cluster without *doen denken aan*;

2: social caused events with Causees acting unintentionally;

3: the *doen denken aan* cluster;

4: physical caused events (source or target domain);

5: social caused events with Causees acting intentionally.

One can see that the best formed and distinct cluster is Cluster 3 with *doen denken aan*, followed by Cluster 2 (social caused events with the Causee acting unintentionally). It is interesting that the physical causation, including both the metaphorical and literal readings, is like a linking element between all other causation types. To some extent this can be explained by the high frequency of the metaphorical expressions with the physical source domain and social or mental target domains (e.g. *een belletje doen rinkelen* “ring a bell”). The social causation with intentional Causees is quite marginal. It is located predominantly at the very top of the map in the area of the least direct causation, although there are also a few observations at the bottom. They represent a small subcluster that emerges after the 20th split with more individuated (definite and singular) and therefore more patient-like Causees than in the upper cluster.

4.4. Lectal variation of *doen*

The analyses in the previous sections were based on the aggregate sample, which included the data from two different geographical varieties and three registers. In this section I focus on the differences between these lects in the use of *doen*. Figure 4.6 shows the normalized frequencies (per million tokens) of *doen* in each lect. One can see that *doen* is less frequent in the Netherlands and in the less formal registers, in full accordance with the previous studies (see Chapter 2, Section 2.3).

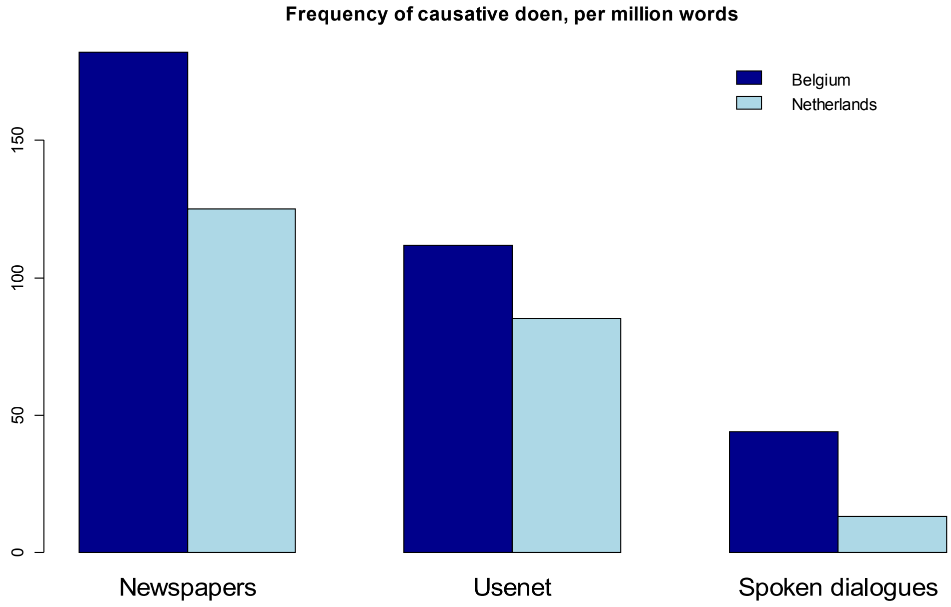


Figure 4.6. Relative frequencies of the causative auxiliary *doen* in six lectal samples.

The average distances between the exemplars in the written registers in Belgium and the Netherlands are similar. They fall very closely at 0.35 for the newspapers, and 0.37 for the Usenet in the two countries. The maximum values are all above 0.70 for these registers. The exemplars found in the spoken data are semantically less diverse: the exemplars of *doen* in the Belgian conversations have the average distance of 0.32 (maximum 0.68), and the observations in the Netherlandic spoken data are even more similar one to another, with the average distance of only 0.24 (maximum only 0.48). One can see the distributions of the distances between exemplars in Figure 4.7. The Netherlandic conversations actually have a bimodal distribution, which shows that there are many exemplars that are located at a very small distance from one another. The Belgian spoken data follow this tendency, but to a lesser extent. On average, the Belgian variants are more semantically diverse than the Netherlandic ones (the mean distances are 0.35 and 0.32, respectively), but this difference is due to the small diversity of *doen* in the Netherlandic conversations.

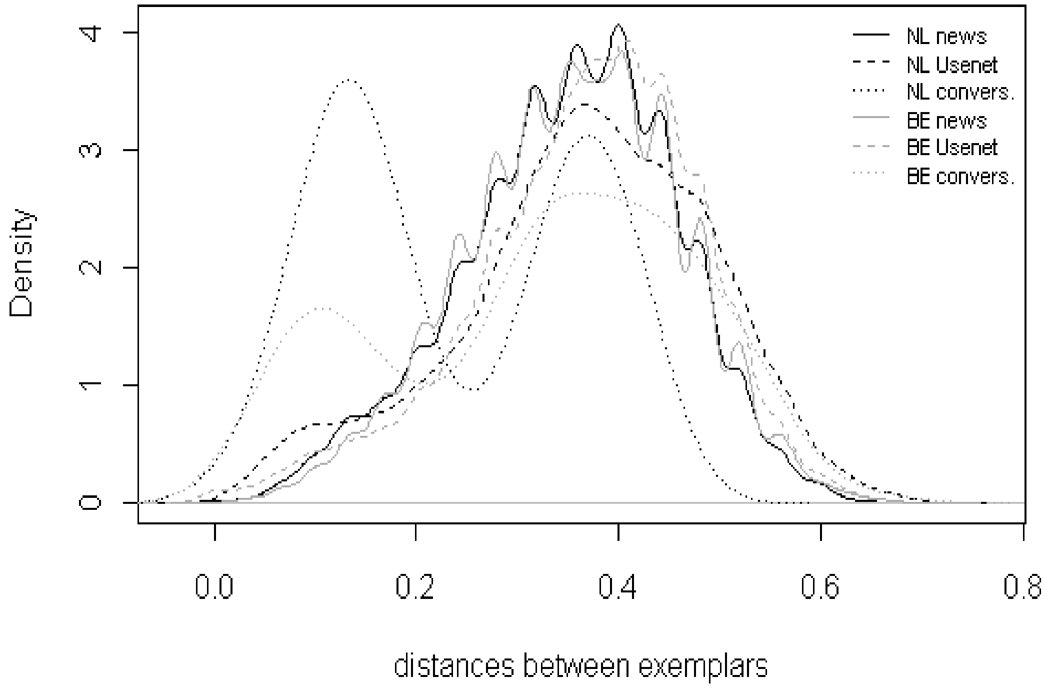


Figure 4.7. Distribution of distances between exemplars of *doen* in six lectal samples.

The next question is whether there is also qualitative difference between the lectal variants of *doen*. To answer it, I plotted the exemplars of the construction in the common semantic space discussed in Section 4.2 according to the subcorpora where they were found.[1] Figure 4.8 shows the positions of the Netherlandic exemplars, and Figure 4.9 displays the distribution of the Belgian *doen*. First of all, one can see that the Netherlandic exemplars display a clearer division between the main senses – mental and non-mental caused events. The Belgian exemplars are distributed more evenly, which suggests that the Belgian *doen* is more monosemous.

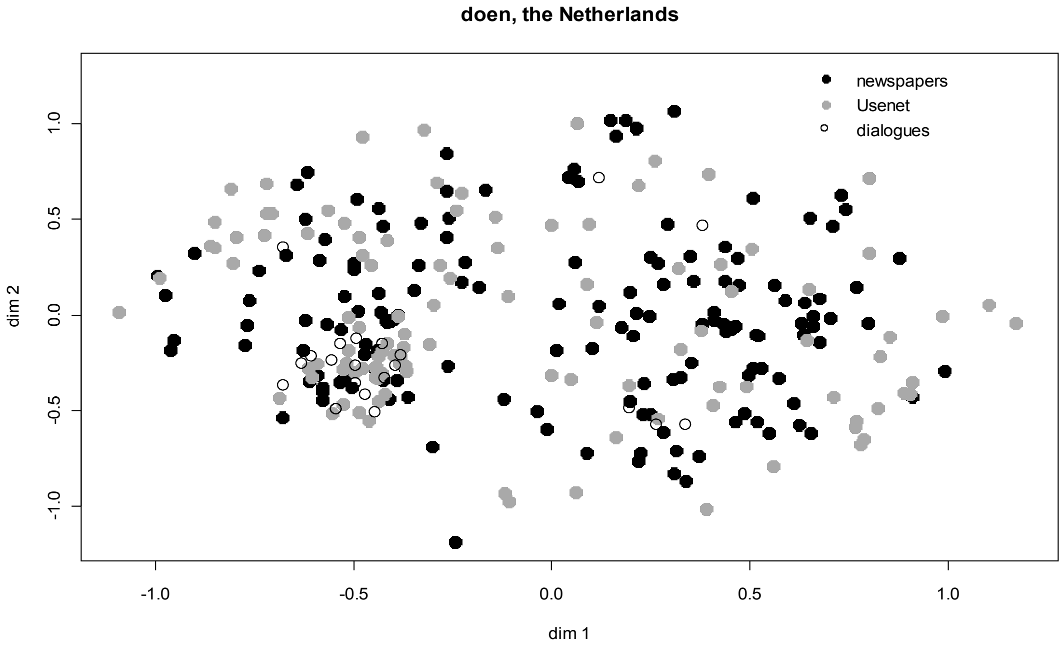


Figure 4.8. Distribution of the Netherlandic exemplars of *doen* in three registers.

One can also see that the Netherlandic exemplars, especially the ones from the spoken data, are more shifted towards the left, 'mental' part of the map, whereas their Belgian and/or written register counterparts are

[1] I have also tried separate MDS solutions for the countries and genres. The resulting semantic spaces are very similar, so I present the exemplars in one common space, which was discussed in detail in Section 4.2.

more evenly distributed. A test of the distribution of the semantic domains of the caused event shows that this difference is statistically significant across the spoken and written registers (Mantel-Haenszel X-squared = 5.3196, d.f. = 1, *p* = 0.021), although separate analyses show that there is no significant geographic difference in the newspapers, and the difference between the country-specific Usenets has borderline significance (*p* = 0.06). Conversely, the spoken data seem to be more 'mental' than the written registers in both countries (Mantel-Haenszel X-squared = 11.7898, d.f. = 1, *p* < 0.001). However, a closer look suggests that this tendency might be due to the highly significant differences in the frequency of *doen denken aan* both between the countries and between the registers. Indeed, after I subtracted the frequency of the collocation from the frequency of all other mental caused events, and applied the same tests, the difference both between the registers and between the countries ceased to be statistically significant at α = 0.05.

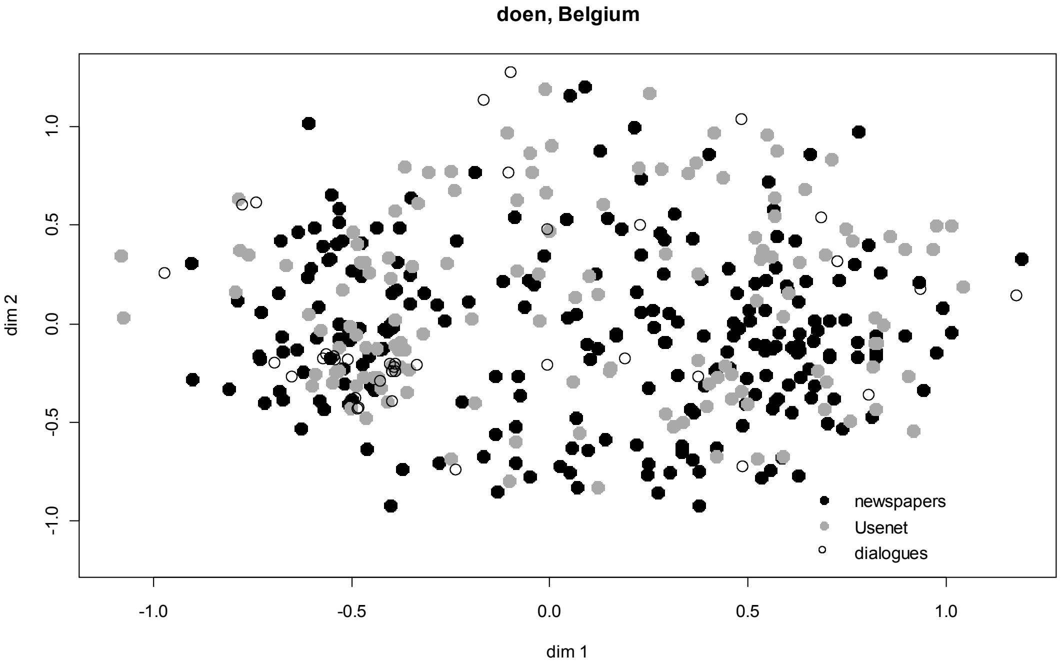


Figure 4.9. Distribution of the Belgian exemplars of *doen* in three registers.

On the basis of the previous studies one could also expect the Belgian variant of *doen* to be less 'direct' with regard to its position on the (in)directness dimension in the semantic map and the corresponding semantic features. Indeed, the exemplars with the highest values on the vertical dimensions come from the Belgian data. The main features associated with indirectness (Causee's intentions and transitivity of the Effected Predicate) are, however, not significantly more frequent in the Belgian sample in general, nor in any of the registers separately.

After studying the general dimensions, one may wonder how the clusters of usage patterns identified in the previous section are distributed lectally. The association plot in Figure 4.10 shows to what extent the observed frequencies of *doen* in a few clusters across the six lects diverge from the expected frequencies. The plot was created with the help of the `vcd` package in R and shows Pearson's residuals in a log-linear model. Vertically, it is organized in five columns, which represent the five large clusters that were shown in the MDS map in Figure 4.5. Horizontally, the plot shows three registers, and two countries within each register. The plot can be interpreted as follows. If the bar is above the line, then the frequency of *doen* in the cluster for the given lect in the given country is higher than one would expect. If it is under the line, then the frequency is too low. The 'taller' the bar, the greater the deviation and thus the more surprising the result. The coloured bars indicate the residuals with the absolute value of 2 (which approximately corresponds to the individual significance of the residual at $\alpha = 0.05$) and 4 ($\alpha = 0.0001$).

Further experiments with different clustering solutions (not shown) do not reveal many significant differences between the countries. Some of them concern the frequencies of specific lexical patterns in very small clusters (e.g. *hoop doet leven* “while there is life there is hope (lit. hope makes live)” is preferred in the Belgian registers, whereas *van zich doen spreken* “get oneself noticed (lit. make speak about oneself)” and *willen* *doen geloven* “want to make believe” are used more frequently than expected by the Netherlandic speakers). The differences between the registers are somewhat more apparent, but they can be explained by the specific communicative functions of the Usenet and spoken data. For instance, reporting the physical reactions as in (26) is typical of international computer-mediated communication in general. This is why the cluster with pronominal Causees and physical caused events is more frequent in the Usenet (especially in Belgium) and is underrepresented in the other registers.

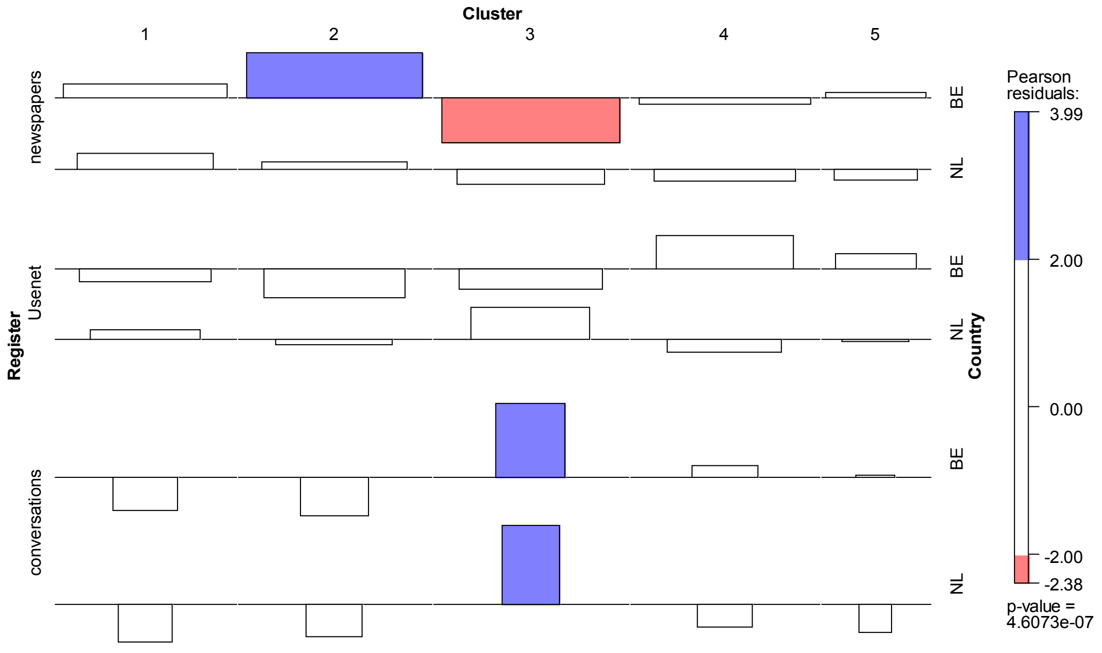


Figure 4.10. Mosaic plot with the proportions of exemplars in five clusters of usage patterns in different lects. The numbers stand for the following clusters. 1: the big mental cluster without *doen denken aan*; 2: social caused events with Causees acting unintentionally; 3: *doen denken aan* cluster; 4: physical caused events; 5: social caused events with Causees acting intentionally.

(26) *En een argument dat me (...) over de grond*
and an argument that me over the ground

zou doen rollen van het lachen.
would do roll from the laughter
“And an argument that would make me (…) roll over the floor laughing.” (*nl.politiek*)

To summarize, there is clear evidence that the lects display quantitative and qualitative variation. As for geographic variation, the Netherlandic *doen* is in general less frequent, less semantically diverse (but only in the spoken data), more fragmented (polysemous) and more associated with mental caused events (in the conversations and to some extent in the Usenet) than the Belgian variant due to the high frequency of *doen denken aan*. The same holds for *doen* in the spoken data in comparison with the written texts, especially the newspapers. In addition, the Belgian newspapers contain much more social caused events with non-volitional Causees (especially quantitative changes along a scale in economy-related articles) than one would expect and significantly fewer constructions *doen denken aan*.

Finally, let us have a look at a very different lect: child language. A search for all occurrences of *doen* in the Dutch CHILDES[1] data has yielded only two contexts, which could be tentatively interpreted as instances of the causative construction with *doen* from the total number of 3000 occurrences of *doen*. The examples are below (the second example can also be interpreted as an attempt to use the resultative construction):

(27) *doe hem daar maar liggen.* [about a toy]
do it/him there but lie
“Make it/him lie there.”

(28) *doet die &m kinderen nat ə worden.* [about rain]

[1] http://childes.psy.cmu.edu/data/Germanic/Dutch/. The data represent transcripts of recordings with Dutch-speaking monolingual children from about 1;10 to 6 years old, but the largest part covers the age from 2 to 3 years.

does those &m children wet ə become

“Makes those children wet.”

This fact supports the conclusion about the marginal status of *doen* in contemporary Dutch. The expression *doen denken aan* “remind of”, highly frequent in the adult speech, is not found in the data. This is not surprising because this kind of metacognitive awareness is beyond children's general cognitive development at the age when the largest part of the data was collected. The same can be said about the abstract social causation. In any case, this subconstruction has to be learnt at a later stage – probably, to a large extent from written sources, especially in the case of the non-mental exemplars.

4.5. Summary

In this chapter I applied different numeric and visualization techniques to explore the semantic structure of *doen*. They yielded the following results.

- The most typical exemplars of *doen* and the most popular features are those associated with affective causation. The least typical exemplars and features correspond to inducive causation, where *laten* would be more appropriate.
- The main dimensions of variation of the exemplars of *doen* are, first, the opposition between mental and non-mental caused events, and, second, (in)directness of causation, which is associated with agentivity of the Causee, reflected in the intentionality of the latter's actions and some other features.
- The most entrenched senses are associated with affective causation, especially the highly frequent collocation *doen denken aan* “remind of” and constructions with social caused events and abstract

Causees, especially quantitative change along a scale.

- The hierarchical network of *doen* reflects the global distinction between the mental and non-mental caused events. At the local levels, it also contains other conceptual, lexical and pragmatic distinctions, such as figurative and quasi-figurative uses of physical Effected Predicates, clusters with various set expressions, and the use of *doen denken aan* in a descriptive 'objective' way.
- There is substantial quantitative and qualitative variation in the use of *doen* in the Netherlands and Belgium and across different registers. *Doen* is more frequent and also more semantically diverse in Belgian Dutch and in the written registers than in the Netherlandic variety and in the spontaneous face-to-face conversations, where it is also restricted to affective causation, mostly expressed with *doen denken aan*. The mental cluster is also more autonomous in the Netherlandic variant.
- According to the child language corpora with approximately 3000 cases of *doen* in various non-causative functions, the causative construction with *doen* is not reproduced at the early stage of acquisition. This implies that it should be learnt at an older age.

Chapter 5. Semantic and lectal variation of the causative construction with *laten*

This chapter continues the semasiological line and presents the analysis of the semantic structure of the causative construction with *laten*. The composition of the chapter mirrors the structure of the previous one. In the beginning, I discuss the family resemblance structure of the category, which is modelled with the help of average distances between the exemplars. Section 5.2 presents the semantic map of *laten* with the most important semantic dimensions. Next, I describe the main usage patterns that emerge as the result of a hierarchical cluster analysis. Section 5.4 examines lectal variation in the use of *laten*. A summary of the results is provided in Section 5.5.

5.1. Intracategorial salience of exemplars of *laten*

As in the previous chapter, I begin with the general facts about the distances between the *laten*-exemplars based on Gower's similarity measure. The average distance between a pair of exemplars was 0.42 (with the standard deviation 0.11), whereas the minimum and the maximum distances were 0 and 0.85, respectively. Both the average and maximum distances were thus greater than the corresponding values for the exemplars of *doen* (see Section 4.1 of the previous chapter), which means that the semantics of *laten* is more diverse.

I also calculated the average distances from every exemplar to the others to find out which exemplars would be the most and the least prototypical according to the family resemblance criterion. As in the case of *doen*, the most typical are the observations that contain the default values of the variables: indicative mood, declarative sentences, predicative function, no modal verbs and negation, no coreferentiality or possession relationships, and an explicit definite Causer. A more interesting finding is that all of the fifty most prototypical exemplars contained human Causers (individuals or organizations). The majority of them were also used in the main clause, referred to unintentional caused events that did not involve any change, like the most prototypical exemplars of *doen*, but, unlike the latter, did not have prepositional complements. Transitivity, the semantic domain of the caused event and the semantic class of the Causee were less prominent features than in the case of *doen*. There were also no specific dominating Effected Predicates, like *denken*. This suggests that the prototypical meaning of *laten* is more abstract than that of *doen* because it has fewer specific features. The exemplar with the greatest number of the most frequent features is (1) with an intransitive Effected Predicate and human explicit nominal Causee – the features that are only slightly more frequent than their alternatives:

(1) *Ook Rombouts laat Van Hecke vallen.*
also Rombouts lets Van Hecke fall
"Also Rombouts abandons Van Hecke." (*De Morgen*, Oct. 2001)

As for the most untypical exemplars, they display a mixture of miscellaneous, less frequent features, which cannot be interpreted as a single usage pattern.

5.2. The exemplar space of *laten*

To model the exemplar space of *laten*, I again used MDS. For visualization purposes I took a randomly selected sample of the size equal to that of the data set with *doen*. The resulting MDS map can be seen in Figure 5.1. Interestingly, the cloud has an empty centre. An explanation, which is in line with the previous findings, is that there are very few features that are shared by most exemplars. Therefore, the exemplars display family resemblance without members that would be highly similar to all others.

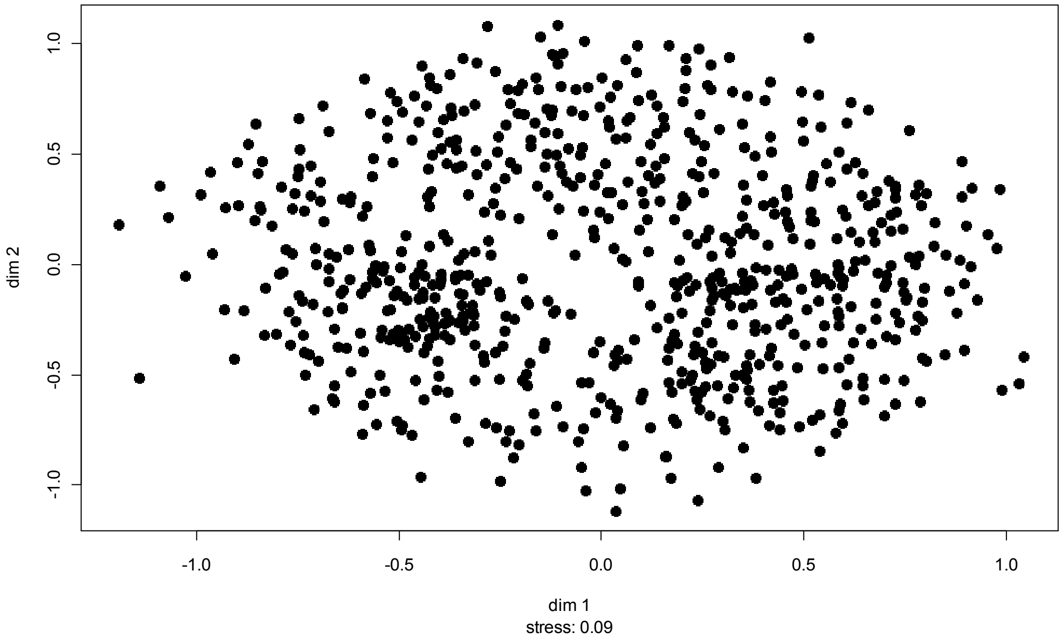


Figure 5.1. MDS representation of the exemplar space of *laten*.

The two dimensions of the solution can be interpreted as follows. The horizontal dimension, as in the case of *doen*, corresponds to the distinction between mental and non-mental caused events (ANOVA *F*-scores for *CausedSemS* and *CausedSemT* are 467.5 on 2 and 716 d.f. and 476.1 on 2 and 712 d.f., respectively, with p < 0.001; the explained variance R^2 is 0.57 for both, which is also the highest observed score). Compare two examples from the extreme left (2) and right (3):

(2) *Laat u niet ontmoedigen en zet uw bijdragen*
let you not discourage and set your contributions

voor *deze* *nieuwsgroep* *gewoon* *voort!*
for this newsgroup as usual forward
“Don't let yourself be discouraged and continue contributing to this newsgroup as usual!” (*nl.financieel.beurs*)

(3) *'s avonds* *willen* *ze* *geen* *bussen* *meer* *laten* *rijden*.
in-the-evening want they no buses more let ride
“They want to cancel buses in the evening.” (*CGN*, fn000795)

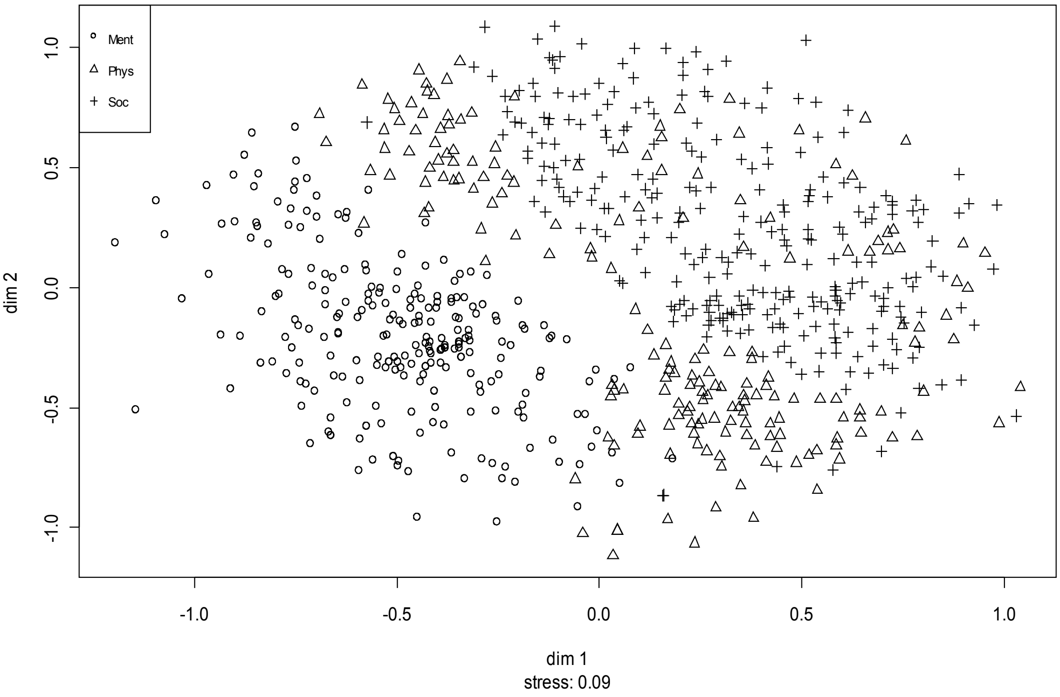


Figure 5.2. Distribution of the source semantic domain of the caused events in the exemplar space of *laten*.

However, these features are also to some extent associated with the vertical dimension, as can be seen from Figure 5.2, which shows that the mental caused events (source domain) are located on average a little lower than the physical and social ones.

The second dimension is associated the most strongly with intentionality of the Causee's actions (unintentional at the bottom, intentional at the top, with a few undefined cases in the middle). The *F*-statistic is 325.1 on 2 and 728 d.f. with $p < 0.001$, and R^2 is 0.47 (again, it is the largest value). Intentionality is followed by the role of the Causee (change of the Causee – no change at all – the Causee causing a change): *F* =191.9 on 2 and 727 d.f. with $p < 0.001$, $R^2 = 0.34$. Both features imply the distinction between patient-like and relatively autonomous Causees. Compare (4), where the implicit Causee is the food that is being prepared, and (5), an exemplar with a Causee who plays an active role. The Causee in (5) is implicit and belongs to the class 'human undefined' because it is not clear whether the action is performed by an individual or organization. The exemplars are located in the extreme top and bottom areas of the map, respectively. A closer inspection shows that the exemplars in the top region very frequently contain human undefined Causees.

(4) *heel e*a heel even heel even laten roerbakken*
very e*a just a little bit just a little bit let stir-fry
heel even en klaar.
just a little bit and ready
“Just a little bit, let (it) stir-fry just a little bit, and it's ready.” (*CGN*, fn000968)

(5) *Is er een reden voor om 4 volmachthouders*
is there a reason for in-order 4 authorized-holders
te hebben? Waarom die andere drie niet laten
to have why those other three not let
schrappen?
drop?
“Is there any reason for having 4 authorized holders? Why not have

the other three dropped?" (*be.finance*)

Thus, the horizontal dimension is the most strongly associated with the semantic domain of the caused event, whereas the vertical dimension is related to (in)directness, as in the case of *doen*. Note that the mental events on average involve less control by the Causee because most of them refer to perception and knowing. However, if one examines the distribution of the variables closely, one will find additional dimensions. The most important one is transitivity of the Effected Predicate, which cuts the map diagonally, as shown in Figure 5.3.[1]

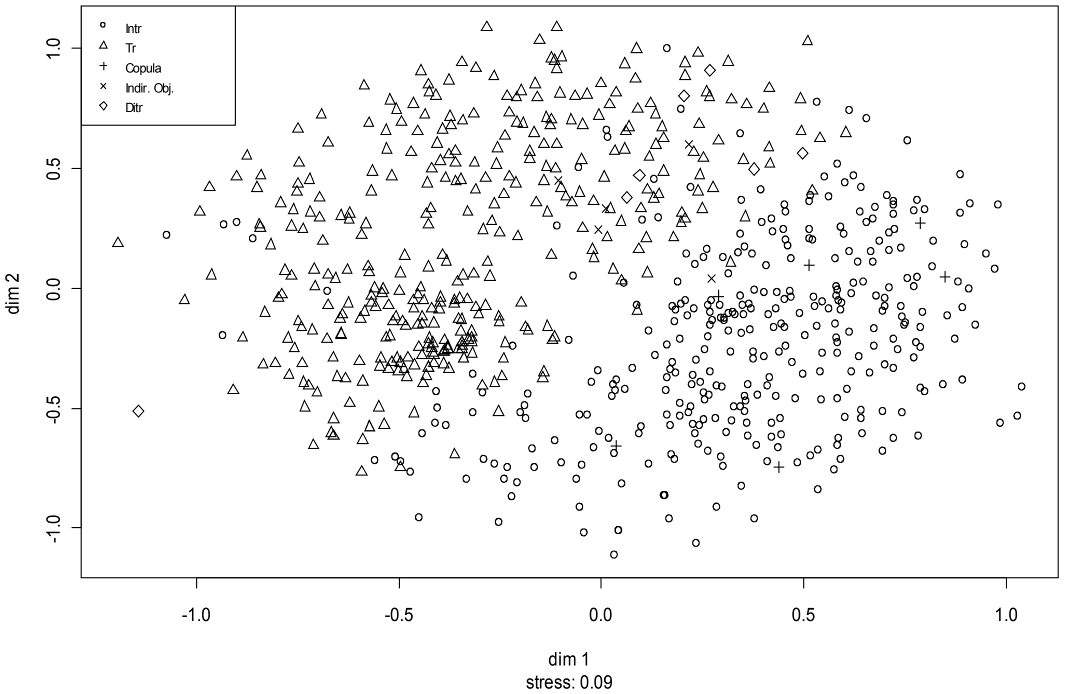


Figure 5.3. Distribution of the Effected Predicate transitivity patterns in the exemplar space of *laten*.

Most intransitives are located in the bottom right part of the map, whereas the transitives populate the top left part. The transitive sector also contains coreferential Causers and Affectees, and most of the implicit and

[1] In the exemplar space of *doen*, a similar cross-cutting effect of transitivity was observed, too, but it was much less outspoken because there were very few transitive Effected Predicates.

prepositionally marked Causees. If one combines this diagonal dimension with the semantic domains of the caused events from Figure 5.2, one can see that the classifications cross-cut the exemplar space. Mental caused events can be subdivided into transitive and intransitive, and so can the physical and social events.

The next step is to examine the density of exemplars. Figure 5.4 shows that the most populated area is located in the left part of the map with mental caused events, as in the case of *doen*. The second densest cluster again corresponds to the non-mental caused events. However, the region at the top with the most indirect causation is more populated than it was on the map with *doen*. A closer examination shows that the densest cluster in the 'mental' part corresponds to several highly frequent expressions (*laten weten* “inform”, *laten zien* “show” and *laten horen* “let hear”). Most exemplars in this cluster relate to providing information, as in (6):

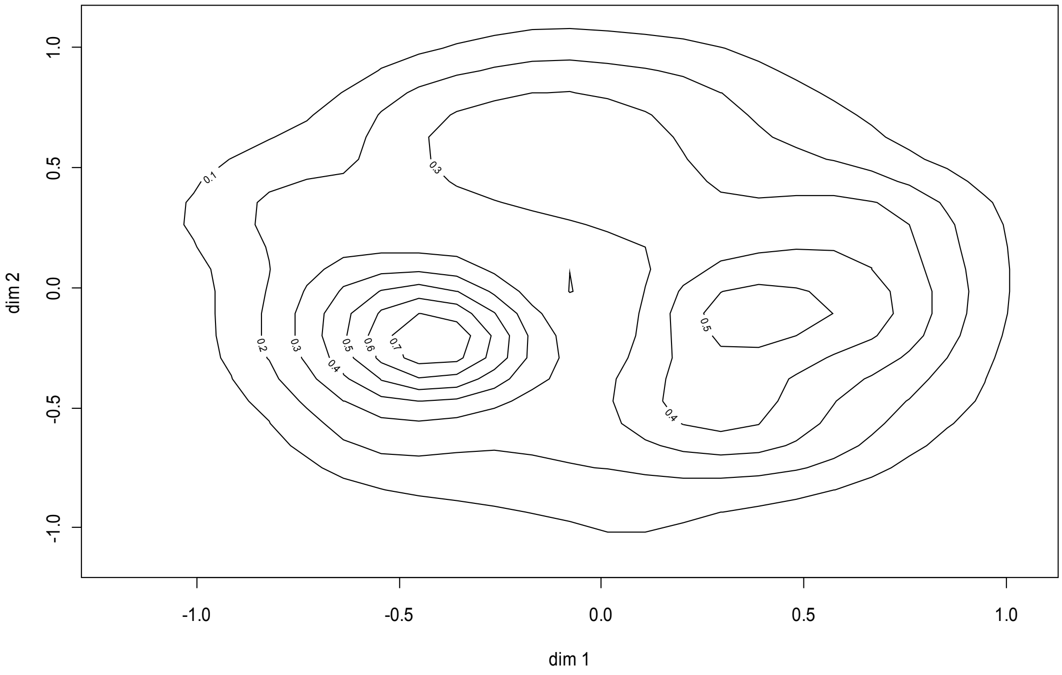


Figure 5.4. Density of the exemplars of *laten*.

(6) *Berlusconi liet gisteren weten de functie van Ruggiero*
Berlusconi let yesterday know the function of Ruggiero
voor zeker zes maanden waar te zullen nemen.
for sure six months true to shall take
"Berlusconi said yesterday that he will fill in the function of Ruggiero for at least six months." (*De Volkskrant*, Jan. 2002)

Note, however, that the area is represented by different Effected Predicates, unlike the area with *doen denken aan*, and the maximal density of the *laten*-exemplars is smaller than that of the *doen*-exemplars.

The densely populated area in the non-mental part contains some frequent fixed expressions, too, e.g. *laten (links) liggen* "ignore" and *laten vallen* "drop, abandon", although these effects are less evident. Therefore, the general conclusion is that the density of exemplars correlates positively with their lexical homogeneity. This suggests that the exemplars with the identical Effected Predicates also tend to be highly similar with regard to other features. The nature and strength of this correlation remain to be explored. It is not clear whether the effect is due to the central role of verbs in the constructional meaning, or if it is a mere artefact of the analysis because the predicate carries a substantial part of the information that is reflected in the other features (e.g. valency, semantic domain of the predicate, reflexivity, etc.).

5.3. Hierarchical cluster analysis of the exemplars of *laten*

The MDS map in the previous section did not suggest a clear cluster structure. The density plot showed a slight indication of three main areas: the distinct densely populated area with the verbs of perception on the left, the area on the right with intransitive Effected Predicates that denote mostly social caused events, and the region with highly agentive Causees at the top, which is rather an extension of the previous area than a separate region. The results of a hierarchical clustering analysis can be seen in Figure 5.5. It does not show a clear picture with a definite optimal number of clusters, either.

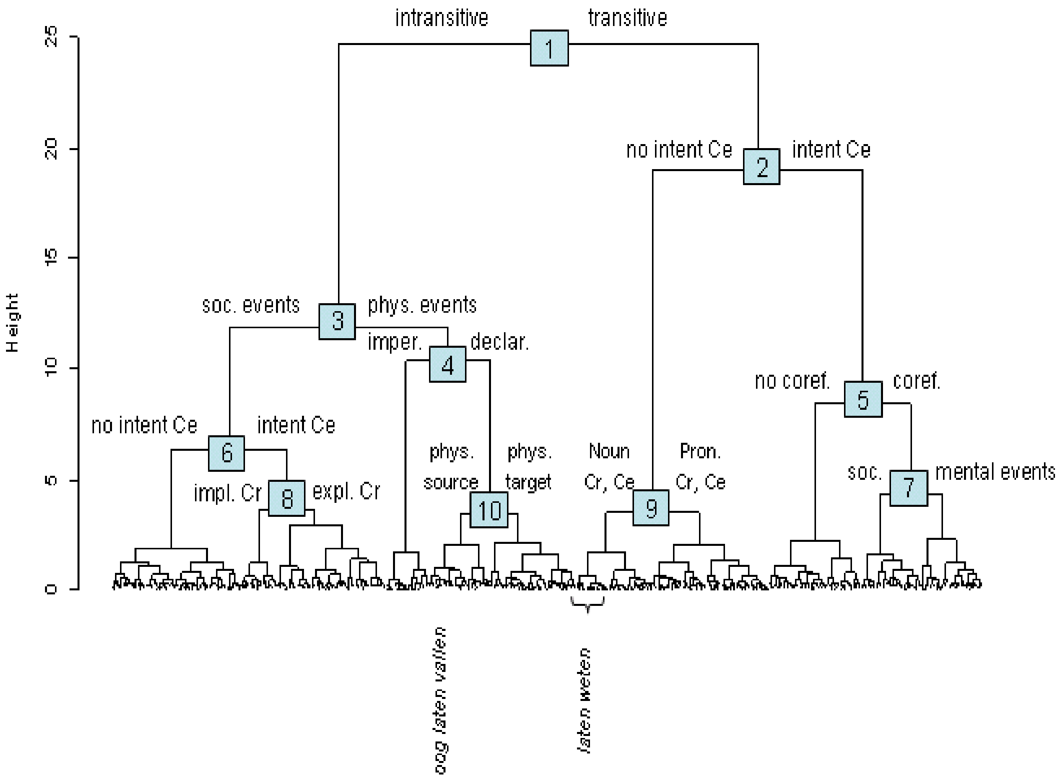


Figure 5.5. Hierarchical clustering of exemplars with *laten*.

As in the previous chapter, I will discuss the 10 top splits (or rather the ten last amalgamations). The first split (1) separates two large groups: intransitive constructions and transitive constructions. According to the length of the 'stems', the intransitive exemplars are more homogeneous than the transitive ones. The transitive cluster bifurcates (Split 2), forming a cluster with Causees acting unintentionally and mostly mental caused events, and a cluster with all other caused events (most of them involve agentive Causees who cause a change intentionally). Note that the cluster with unintentionally acting Causees is more homogeneous than the cluster with intentionally acting ones (the same holds also for Split 6 of the social caused events). This can be explained by high similarity between the exemplars with unintentionally acting Causees. These exemplars are mainly represented by *laten weten* “inform”, *laten zien* “show” and similar collocations. The 'unintentional' cluster divides (Split 9), in its turn, into a cluster with nominal Causers and Causees and mostly clausal Affectees, and a cluster with pronominal Causers and Causees and predominantly non-clausal Affectees. The former refers to providing information, and is more 'intellectual', e.g. (7), whereas the latter involves showing something to somebody and more frequently refers to physical perception, e.g. (8).

(7) *Berlusconi liet gisteren weten de functie van Ruggiero*
Berlusconi let yesterday know the function of Ruggiero
voor zeker zes maanden waar te zullen nemen.
for sure six months true to shall take
“Berlusconi said yesterday that he will fill in the function of Ruggiero for at least six months.” (*De Volkskrant*, Jan. 2002)

(8) *Ik liet een vriendin een song horen.*
I let a friend a song hear
“I let a friend hear a song.” (*De Standaard*, Oct. 2001)

The cluster with non-mental caused events and transitive Effected Predicates, in its turn, bifurcates in Split 5, forming a cluster with coreferential Causers and Affectees, and a cluster with non-coreferential contexts. The latter are represented by different types of human interaction, for example, the service frame, as in (9):

(9) *Dit document kunt u bekomen op uw postkantoor*
this document can you receive on your postoffice

of laten opsturen door de Internationale dienst.
or let send by the International service
“You can receive this document at your post office or have it sent by the International service.” (*be.finance*)

The coreferential contexts (Split 7) are subdivided according to the domain of the caused events (source and target): social or mental. Compare (10), where the Affectee (i.e. the Causer) is affected as a social entity, and (11), which denotes mental impact:

(10) *Alsof een groot blok als de CD&V zich zomaar laat*
as if a big block as the CD&V self just lets
gijselen door haar kleine kartelpartner.
hold-hostage by its small cartel-partner
“As if such a large block as the CD&V can be held hostage by its small cartel partner.” (*be.politics*)

(11) *Je moet je niet door cijfers laten intimideren.*
you must you not by numbers let intimidate
“You shouldn't let yourself be intimidated by numbers.” (*nl.beurs*)

As the examples show, most of the Effected Predicates in these clusters specify some negative influence (cf. Loewenthal 2003). The clusters also contain constructions with the middle voice semantics:

(12) *Welke gevolgen deze ontwikkeling heeft voor de westerse*
what consequences this development has for the western
economie laat zich raden.
economy lets itself guess
“It's easy to guess what consequences this development will have for

the western economy.” (*be.finance*)

(13) *Hij vertegenwoordigde een ander Duitsland, dat zich*
he represented a different Germany which itself
niet zo gemakkelijk liet exporteren (...)
not so easily let export
“He represented another Germany, which could not be easily exported (...).” (*De Standaard*, Jan. 2002)

The large intransitive cluster from the top split branches in Split 3, forming a cluster with social caused events and a cluster with non-social ones (mostly physical caused events, in both literal and figurative readings). The non-social cluster then divides in Split 4 into a cluster with mostly imperative contexts, e.g. (14), and the rest. The contexts are intersubjective and informal. The imperative form of the construction *laat* is often followed by discourse particles *maar*, *eens*, *even*:

(14) *laat maar liggen joh.*
let but lie joh
“Leave it where it is, hey.” (*CGN*, fn007962)

The rest of the non-social cluster is split into a cluster with physical target events and a cluster with figurative and idiomatic expressions. The cluster with the physical target events contains caused events that involve different physical forces and processes that enable the Causer to bring about the caused event, such as gravitation, e.g. (15), inertia, energy exchange between a physical body and the environment, as well as the energy of tools and mechanisms, e.g. (16). There are also a number of contexts where the Causer simply leaves an object or person unattended and thus causes a damage, e.g. (17).

(15) *Dreigend draaien ze wijde cirkels en laten hun*
ominously turn they wide circles and let their
dodelijke lading vallen op de heuveltoppen (...)
deadly cargo fall on the hill-tops
"Ominously, they make wide circles and drop their deadly cargo on the hill tops." (*AD*, Nov. 2001)

(16) *'s avonds willen ze geen bussen meer laten*
in-the-evening want they no buses more let
rijden.
ride
"They want to cancel buses in the evening." (*CGN*, fn000795)

(17) *Ze willen dan achteraf de dader lynchen, kielhalen,*
they want then afterward the criminal lynch keelhaul
laten verhongeren etc...
let starve etc...
"Afterward they want to lynch, keelhaul or starve the criminal to death." (*nl.politiek*)

The idiomatic cluster contains many figurative expressions that involve body parts, e.g. *zijn oren laten hangen naar iemand* "listen to someone (lit. let one's ears hang to someone)", *zijn handen laten wapperen* "get to work (lit. let one's hands wave)", *zijn oog laten vallen op iets/iemand* "spot, let one's eye fall on something/someone", as in (18):

(18) *Anderlecht heeft zijn oog laten vallen op Mahan*
Anderlecht has his eye let fall on Mahan
Mondakan.

Mondakan

"Anderlecht let its eye fall on Mahan Mondakan." (*De Standaard,* Feb. 2002)

The figurative expressions of this kind are dramatically different from the idiomatic expressions with *doen*: in most of these *laten*-contexts, there is no metaphorical transfer of energy from the Causer to the Causee because they form one entity. The role of the Causer, who is no longer an external source of energy, is reduced to being responsible for the caused event. This also holds for most non-figurative uses of *laten*.

Finally, the large cluster with social caused events formed by Split 3 divides into clusters with unintentional and intentional behaviour of the Causee. The majority of the Causees acting unintentionally are abstract entities, especially events, as in (19):

(19) *De Vlaamse partijen mogen dit niet laten gebeuren.*
the Flemish parties may this not let happen
"The Flemish parties may not let this happen." (*be.politics*)

The cluster with intentional Causees contains a small subcluster with a large proportion of implicit Causers, often in infinitival complements, and a cluster with all others (Split 8). The former cluster contains the cases when it is not clear or important who exactly is responsible for the causation, as in (20):

(20) *Bij de fusie is het politieke besluit genomen*
by the coalitionis the political decision taken
om beide voorzitters te laten gaan.
in-order both chairpersons to let go
"During the process of creating the coalition, a political decision has

been made that both chairpersons should leave." (*AD*, Apr. 2002)

Note that the example contains the passive construction and the Causer is not mentioned anywhere. The focus is on the caused event, which is to take place in the future. Many contexts in the cluster describe plans, intentions, goals, decisions or wishes. On the contrary, the cluster with the explicit Causers frequently contains contexts with undesirable and unplanned caused events that the Causer is blamed for:

(21) *Manager Taillieu verkwanselde deze wedstrijd door*
manager Taillieu squandered this match by
centrale middenvelders naar de vleugels te laten spelen.
central midfield-players to the wings to let play
"Manager Taillieu squandered this match by letting central midfield players play down the wings." (*be.sport.football*)

To summarize, the results of this cluster analysis demonstrate the following. The top-level clusters are based on the schematic constructional patterns (transitivity). Recall that transitivity represents a fundamental syntactic and conceptual distinction for analytic causatives in all languages, according to Kemmer and Verhagen's (1994) study. The semantic domain of the caused event and the semantic classes of the participants come into play later. The mood and coreferentiality play a more prominent role than in the cluster solution of *doen*. The figurative vs. literal distinction is observed for physical caused events, as was the case with the previous construction. Lexical similarity is much less important than in the case of *doen*: there are very few distinct lexically specific clusters. Interestingly, the features of the Causer play a more important role here: nominal or pronominal (together with the Causee), and implicit or explicit. The distinction between the imperative and declarative mood also implies the Causer's absence or presence. These distinctions reflect different roles of human Causers in bringing about the event, as well as different communicative functions of the message.

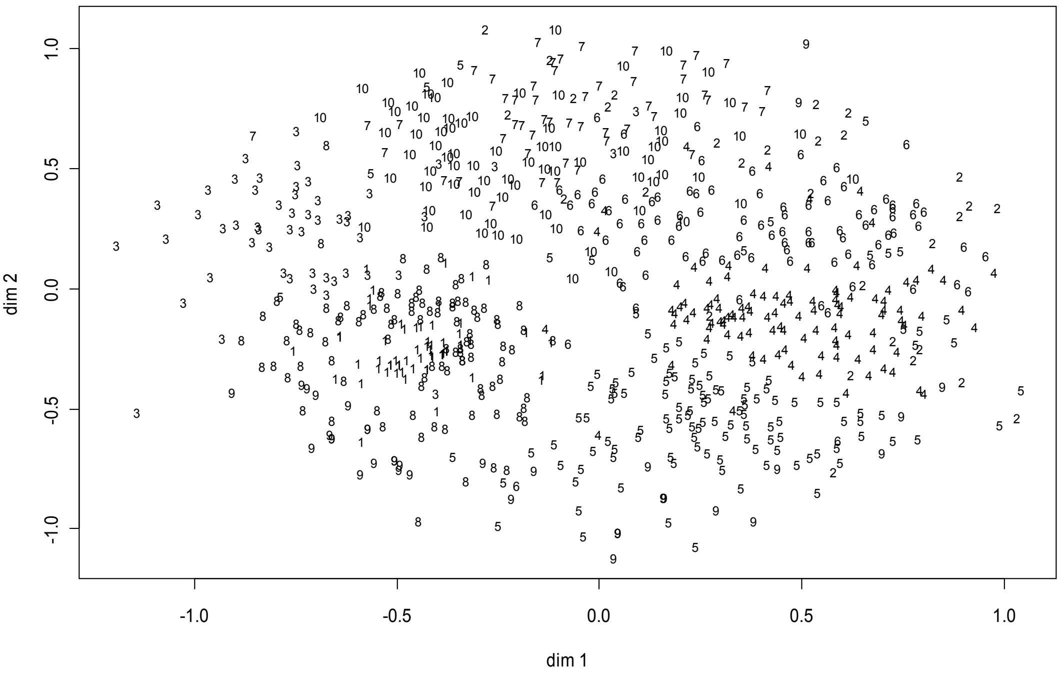


Figure 5.6. Ten usage clusters of *laten* on the MDS map.

To give an idea how the clusters are related to the dimensions of the MDS solution from Section 5.2, I projected ten clusters of usage patterns on the MDS map. The patterns are represented as numbers in the following order (the numbers in the brackets refer to the corresponding examples in the previous text):

1: transitive Effected Predicates, mental caused events, and providing information (7);

2: intransitive Effected Predicates, social caused events, Causees acting intentionally, and implicit Causers (20);

3: transitive Effected Predicates, Causees acting intentionally, mental caused events, and coreferential Causers and Affectees (11);

4: intransitive Effected Predicates, social caused events, Causees acting unintentionally (19);

5: intransitive Effected Predicates, physical caused events (figurative and literal), explicit Causers (15) – (18);

6: intransitive Effected Predicates, social caused events, Causees acting intentionally, explicit Causers (21);

7: transitive Effected Predicates, Causees acting intentionally, coreferential Causer and Affectee, social caused events (10);

8: transitive Effected Predicates, mental caused events, and pronominal Causer and Causees, mostly physical perception (8);

9: intransitive Effected Predicates, non-social caused events and imperative sentences (14);

10: transitive Effected Predicates, Causees acting intentionally, no coreferentiality, mostly physical caused events (9).

5.4. Lectal variation of *laten*

From the quantitative perspective, the construction displays less lectal variation than *doen*. The bar plot in Figure 5.7 shows that the geographic difference is outspoken only within the newspaper register. The Netherlandic newspapers contain more occurrences of *laten* than the Belgian ones (per million tokens). As in the case of *doen*, the construction occurs less frequently in the spoken data than in the written texts, although the difference is less dramatic.

Measuring the distances between the exemplars in the six lects reveals no dramatic national differences between the average distances in the Netherlandic and Belgian registers, although the between-register differences are quite outspoken: the smallest average distances are between the exemplars in the Netherlandic and Belgian newspapers (0.38 for both countries), followed by the conversations (0.39 and 0.40, respectively) and finally the Usenet (0.43 and 0.45). This relatively high density of the newspaper exemplars may be due to the frequently used fixed collocations like *laten weten* “inform”. It is worth mentioning that all maximum values of the distances are larger in the Netherlandic registers than in the corresponding Belgian ones (0.80 vs. 0.74 in the newspapers, 0.85 vs. 0.83 in the Usenet, and 0.83 vs. 0.73 in the conversations). This means that the Netherlandic construction is more tolerant to untypical 'outliers' than the Belgian variant.

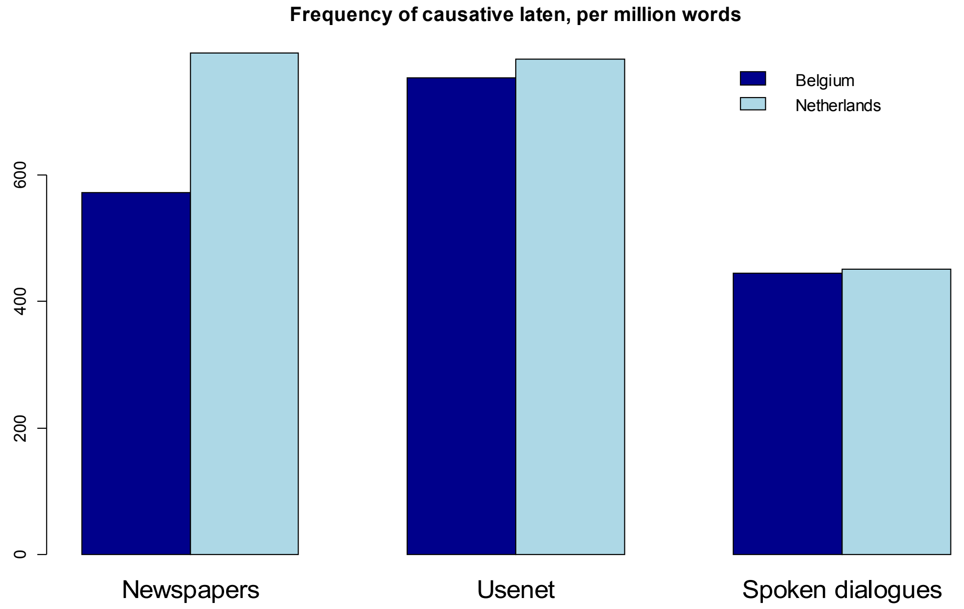


Figure 5.7. Relative frequencies of *laten* (per million tokens) in six lectal samples.

Figure 5.8 shows the distribution of the distances in the six lects. The curves display great similarity, although the small 'hump' in the left-hand tail of the Netherlandic newspaper curve suggests a densely populated compact cluster.

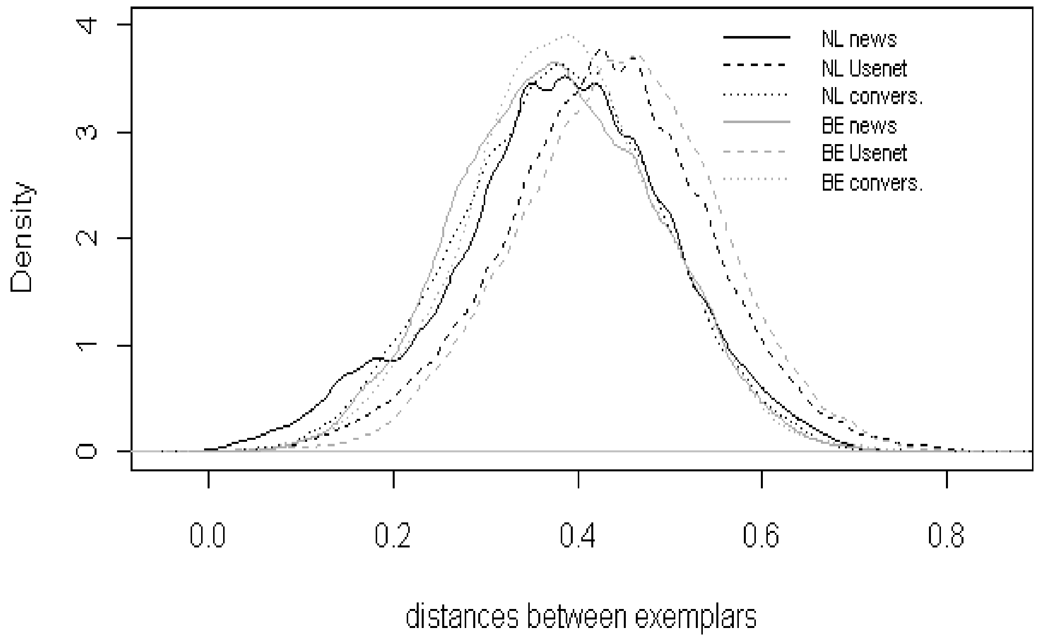


Figure 5.8. Distribution of distances between exemplars of *laten* in six lectal samples.

The next question is whether there are any differences between the lectal samples with regard to the semantic dimensions. To explore this, I plotted the lectally specific exemplars on the common semantic space described in Section 5.2. The results are shown in Figures 5.9 (three registers of the Netherlandic data) and 5.10 (the registers of the Belgian subsample). Again, the Netherlandic variant seems to have a sharper separation between the mental and non-mental caused events. It also forms tighter clusters, especially the one in the mental part. As in the case of *doen*, the Belgian exemplars display fewer clustering patterns and are distributed more homogeneously.

As the MDS maps show, the Netherlandic *laten* has a more densely populated mental causation part than the Belgian variant. A chi-squared test of the frequency ratios of the mental and non-mental caused events (target domains) in the entire Netherlandic and Belgian samples shows that this tendency is significant across the three registers (Mantel-Haenszel X-squared = 53.3904, d.f. = 1, p < 0.001). However, one can see that this difference is to a large extent due to the cluster associated with the sense of giving information or showing, and with specific collocations (*laten zien* "show", *laten horen* "let hear", *laten weten* "inform", etc.). This cluster is more dense in the Netherlandic than in the Belgian data. In fact, the Belgian sample seems to have even slightly more mental caused events than the Netherlandic sample when the observations with these three verbs are removed (Mantel-Haenszel X-squared = 4.7992, d.f. = 1, *p* = 0.028). These Effected Predicates are used the most often in the Netherlandic newspapers (33% of all exemplars), and the least frequently in the Belgian ones (only 4%). Therefore, the difference in the frequencies of these verbs might explain the large quantitative difference in the relative frequencies of *laten* between the Netherlandic and Belgian newspapers (Figure 5.7).

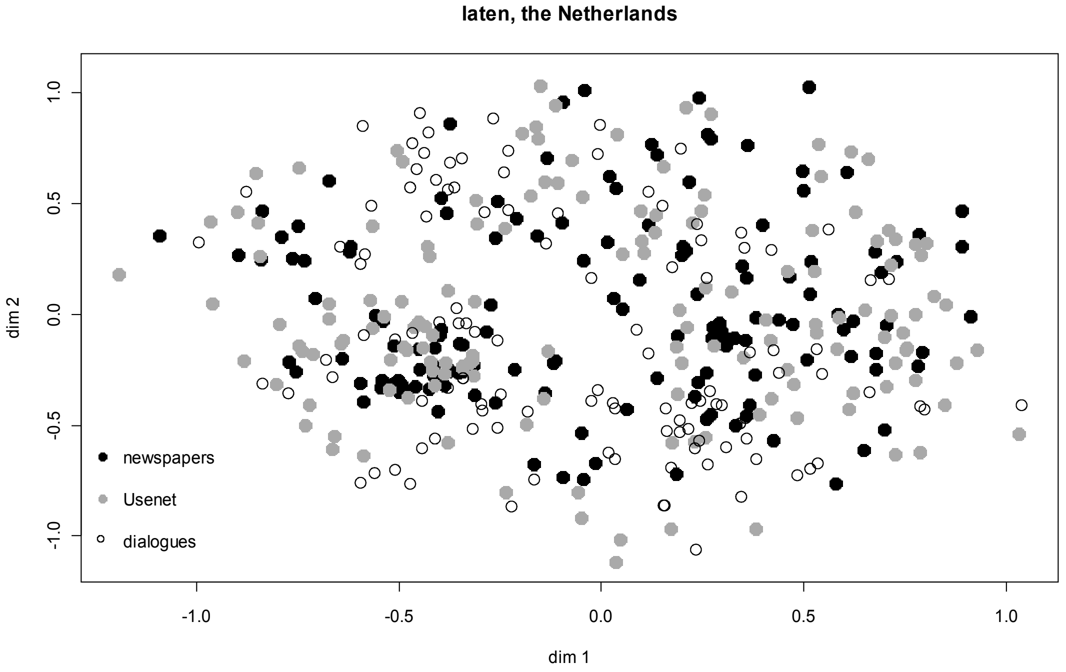


Figure 5.9. Distribution of the Netherlandic exemplars of *laten* in three registers.

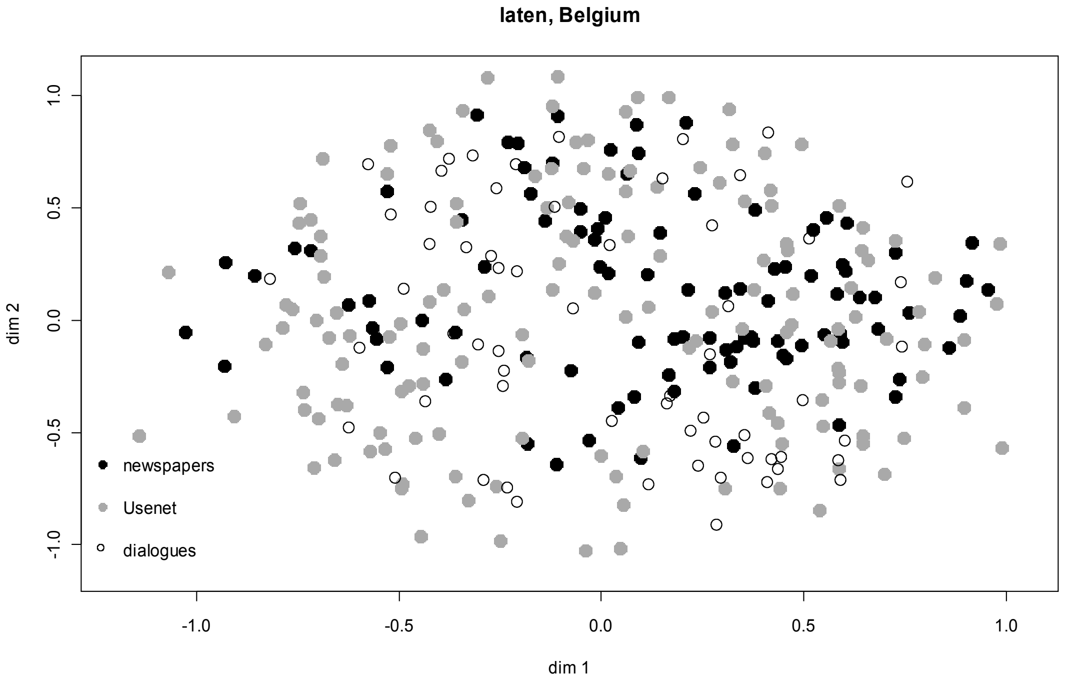


Figure 5.10. Distribution of the Belgian exemplars of *laten* in three registers.

As for the vertical dimension, there seems to be variation, too. A statistical test with the registers as strata shows that the Netherlandic data contain significantly more unintentionally acting Causees than the Belgian sample, especially in the newspaper register (Cochran-Mantel-Haenszel $M^2 = 8.1808$, d.f. = 2, $p = 0.017$). A deeper investigation demonstrates that this preference is due to the same lexical effects because perception and knowledge denoted by the above-mentioned three Effected Predicates are normally uncontrollable mental processes.

Let us now zoom in on the specific semantic areas. The association plot in Figure 5.11 demonstrates how ten usage clusters, which were shown in Figure 5.6, are represented by the lects. The cluster numbers above the columns stand for the usage clusters described in the previous section. The association plot shows very strong deviations from the expected frequencies for Cluster 1 and Cluster 5. The Netherlandic newspapers contain a very large number of exemplars that belong to Cluster 1, which is mainly represented by *laten weten*. Note also that

Cluster 8, which refers to showing, is overrepresented in the Netherlandic conversations. Cluster 5 (intransitive Effected Predicates, physical causation and explicit Causers) is also strongly overrepresented in the Netherlandic spoken data.

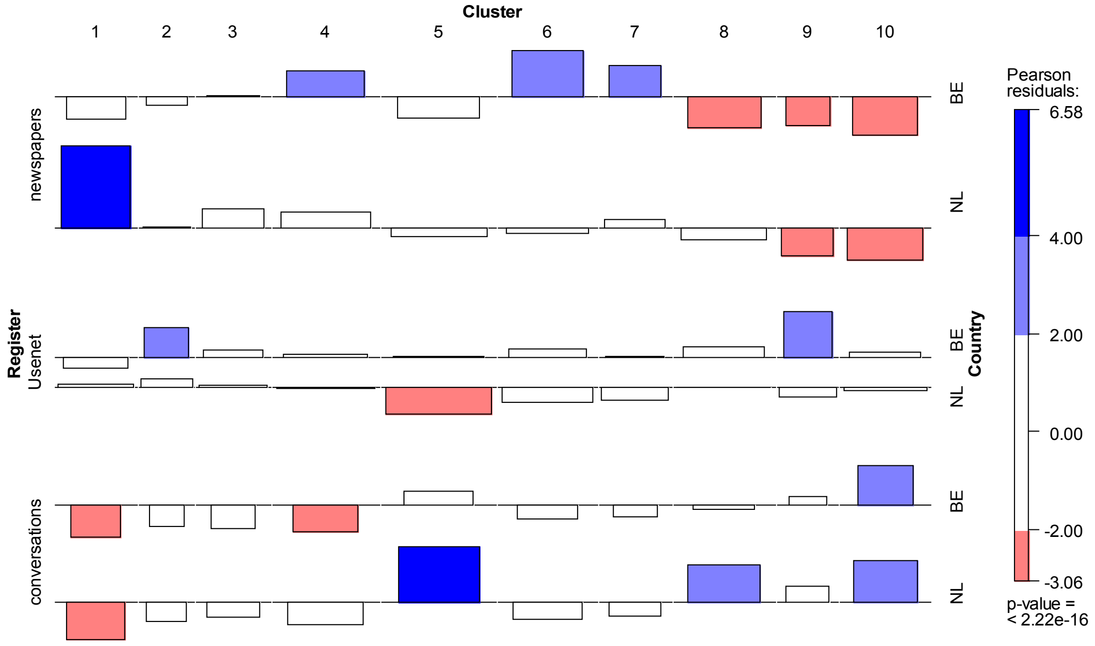


Figure 5.11. Lectal distribution of ten clusters of *laten*. 1: transitive Effected Predicates, Causees acting unintentionally (e.g. *laten weten*); 2: intransitive Effected Predicates, social caused events, Causees acting intentionally, implicit Causers; 3: transitive Effected Predicates, Causees acting intentionally, mental caused events, coreferential Causers and Affectees; 4: intransitive Effected Predicates, social caused events, Causees acting unintentionally; 5: intransitive Effected Predicates, physical caused events, explicit Causers; 6: intransitive Effected Predicates, social caused events, Causees acting intentionally, explicit Causers; 7: transitive Effected Predicates, Causees acting intentionally, coreferential Causer and Affectee, social caused events; 8: transitive Effected Predicates, Causees acting unintentionally, and pronominal Causer and Causees (mostly verbs of perception); 9: intransitive Effected Predicates, non-social caused events and imperative sentences; 10: transitive Effected Predicates, Causees acting intentionally, no coreferentiality, mostly physical caused events.

Some of the more moderate deviations are found in Clusters 4 and 6 with intransitive constructions and social caused events. As the similar sense of *doen*, this type of causation is overrepresented in the Belgian newspapers. This might serve as evidence of the more 'objective' and 'social' orientation of this Belgian register. The 'social' Cluster 7 with coreferential Causers and Affectees is also highly frequent there. Cluster 10 with many physical events (service frame) is overrepresented in the spoken data in both countries – probably because this sense is very common when people talk about everyday activities. Thus, a great part of the lectal variation comes from the differences in the referential contents and communicative functions of the lect.

To summarize, the most prominent differences involve the senses of informing and showing, which are more frequent in the Netherlandic lectal samples, especially in the newspapers. Interestingly, the Belgian newspapers again demonstrate preference for more 'objective' social caused events in different contexts, and seem to disfavour the main 'subjective' senses.

Finally, the data from the Dutch component of CHILDES suggest that the first occurrences of *laten* in children's speech are in the combination with *zien.* It is also the most frequent *laten*-construction in the entire corpus (36 from 91). However, very soon it is followed by subconstructions with non-mental Effected Predicates, such as *staan* “stand” (13 occurrences), *vallen* “fall”(9), *liggen* “lie”(5), *doen* “do” (3), *rijden* “ride, drive” (3), *zitten* “sit” (3), *zwemmen* “swim” (3) and several less frequent ones. There are also occurrences of the mental *laten horen* “let hear” (5) and *laten kijken* “let look (at)” (4). Thus, the non-mental sense has a higher type frequency, but the most frequent type is *laten zien*. These tendencies are very much in line with the patterns found in the adult corpora that are studied here. The *laten weten* construction, reflexives and transitive effected predicates with explicit Affectees, do not occur –

probably because of their specific pragmatic functions and complex structure. This suggests that the physical caused events and physical perception are the most basic senses of the construction from the ontogenetic point of view.

5.5. Summary

In this chapter, an intracategorial analysis of the exemplar space of *laten* has revealed the following results.

- In general, the semantic structure of *laten* is more heterogeneous than that of *doen* because there are more conceptual dimensions of variation, and the distances between the exemplars are on average greater. The exemplars are more evenly distributed in the space, which results in less prominent differences in the density between semantic regions.
- The semantic space of *laten* is organized alongside the dimensions of mental vs. non-mental caused events and direct vs. indirect causation, which is primarily associated with intentionality of the effected event on the part of the Causee. The same dimensions were also found in the case of *doen*. However, some features – most importantly, transitivity – contribute to the structure of the *laten*-space in a more obvious way.
- Like in the case of *doen*, the most densely populated areas on the map involve lexicalized prefabs, especially *laten zien* "show, let see" and *laten weten* "inform, let know".
- Similar distinctions have been found in a cluster analysis of the *laten*-exemplars. The most general opposition is that between transitive and intransitive constructions. The other distinctions involve the semantic domain of the caused event, the intentionality

of the Causee's actions, coreferentiality, clause mood and different properties of the Causer. The analyses have also shown that the lexical similarity plays a less important role in the clustering in comparison with *doen*.

- The construction is subject to relatively modest geographic variation in terms of frequencies. The 'mental' cluster in the Netherlandic variant, as in the case of *doen*, displays greater autonomy and collocational effects related to the expressions of providing information (*laten weten* “inform”, which is especially frequent in the newspaper register) and demonstrating something to the Causee (*laten zien* “show, let see”). As in the case of *doen*, the usage patterns associated with more 'objective' social causation were found to be more popular in the Belgian newspapers than expected. Again, the Belgian journalists seem to shun the more 'subjective' uses. There is also variation in the fine-grained meanings expressed by different registers, which can be explained by different communicative functions and referential situations associated with the lects.
- The child language data exhibits frequency patterns that are similar to the use of the construction by the adults, with the exception of the more formal and complex structures. This shows that the construction is successfully learnt at an early age.

Chapter 6. *doen* vs. *laten*: division of semantic and lectal labour

According to an authoritative reference grammar of Dutch *Algemene Nederlandse Spraakkunst*, the differences between the use of *doen* and that of *laten* are difficult to pinpoint. The variation involves a complex interplay of semantic, regional and stylistic factors (ANS 1997: 1015). This chapter is an attempt to disentangle these sources of variation. It begins with an exploration of the distance matrix with exemplars of *doen* and *laten*, and discusses the members with the highest and lowest cue validity with regard to the category they belong to. Section 6.2 focuses on the regions of conceptual overlap of *doen* and *laten* and their distinctive areas as represented in an MDS map. Section 6.3 describes the distinctive features of the two constructions and tests them in a series of confirmatory analyses. In Section 6.4 I explore the lectal differences in the division of labour between the two constructions in two national varieties and three registers of Dutch. The final Section 6.5 summarizes the findings.

6.1. Cue validity of exemplars of *doen* and *laten*

As in the previous chapters, the first step in the analysis involved creating a matrix of distances between the exemplars. This time, the matrix included the exemplars of both constructions. On the basis of the matrix, I calculated the cue validity of every exemplar with regard to the category it belonged to (*doen* or *laten*). To do so, I took the average distance from the exemplar to all members of its own category and divided it by the average distance from the exemplar to all exemplars of both categories. The exemplars of *doen* with the highest cue validity (the lowest ratio of the average distances) all referred to affective mental association causation with the fixed expression *doen denken aan* “remind of”. The *doen*-exemplars with the lowest cue validity (the highest ratio of the distances) referred mostly to inducive causation, containing non-mental caused events, implicit human undefined Causees acting intentionally and transitive Effected Predicates. The example below had the lowest cue validity, according to this operationalization:

(1) *(...) 'k heb*z ze opnieuw doen vullen want dat*
I have them again do fill because that
waren gevulde tanden.
were filled teeth
“(...) I had them filled again because those were filled teeth.”
(*CGN*, fv400656)

As far as *laten* is concerned, the highest cue validity is found for the *laten*-exemplars which are in fact very similar to the exemplars with the lowest cue validity of *doen*: transitive Effected Predicates, human undefined intentional Causees and physical caused events, as in the following example:

(2) *deze heb ik dus bij de Hema laten afdrukken .*
these have I thus at the Hema let print
“I've had these printed at Hema.” (*CGN*, fn007839)

The exemplars that were the least distinctive of *laten* contained abstract Causers, intransitive Effected Predicates and explicit nominal Causees. Apart from that, the examples are very diverse. Below is one of them:

(3) *'Als de Verenigde Staten maatregelen nemen*
if the United States measures take
die de uitvoer van Europees staal laten teruglopen, dan
that the export of European steel let go-back then
zijn we gerechtigd tegenmaatregelen te nemen',
are we entitled countermeasures to take
aldus Lamy.
according-to Lamy
"If the United States takes measures that reduce the export of European steel, we are entitled to take countermeasures," said Lamy. (*De Volkskrant*, March 2002)

As has been mentioned, the *doen*-exemplars with the lowest cue validity scores with regard to the category are very similar to the ones with the highest cue validity of *laten.* However, the reverse does not hold. This might be explained by the strong exemplar effects, which keep the extremely dense cluster with *doen denken aan* pure from *laten*.

Another observation is that the cue validity scores for *doen* are very similar to the family resemblance scores discussed in the corresponding semasiological study, whereas for *laten* this is not so. Section 7.1 in Chapter 7 treats this issue in detail and offers an explanation.

6.2. The common exemplar space of *doen* and *laten*

The next step of the onomasiological analysis was to construct the common exemplar space of the two constructions in order to identify the areas of their overlap. Because the entire data set was too large for computation and interpretation, I took the random sample of *laten* that was discussed in Section 5.1 with 731 observations and 106 randomly selected exemplars of *doen*, so that the general ratio of *doen* and *laten* in the corpus should be preserved. The resulting MDS map is shown in Figure 6.1.

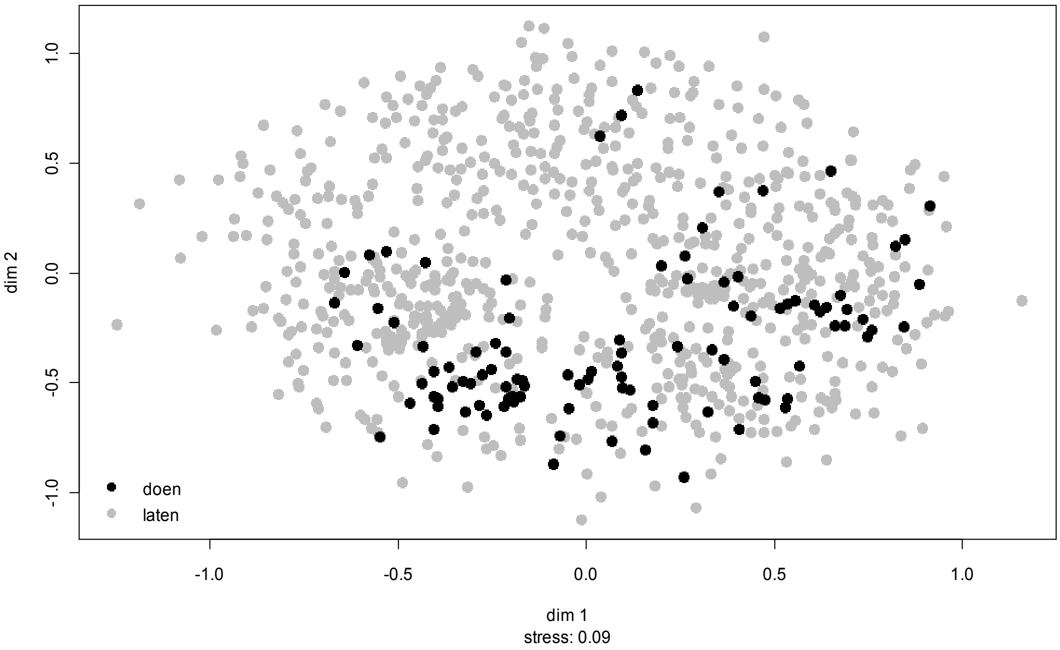


Figure 6.1. Exemplars of *doen* and *laten* in a common space.

The space is dominated by the *laten*-exemplars due to their much greater frequency in the sample, and therefore strongly resembles the semasiological map of *laten* in the previous chapter. Still, a closer inspection shows that the relative positions of the *doen*-exemplars with regard to one another are also very similar to those observed in the semasiological analysis of *doen*. The cluster with *doen denken aan* can be found on the left, in the 'mental' part, although somewhat lower than its old position, whereas the social causation patterns can be found on the right. The most important difference is that the semantic regions and exemplars of *doen* have slightly shifted anticlockwise. This is due to the different structure of the *laten* space. Recall that the latter, as was mentioned in the previous chapter (Section 5.2), is cut diagonally by transitivity of the Effected Predicates, with the intransitives in the bottom and right part, and the transitives in the top and left sector. Because most exemplars of *doen* are intransitive, they are concentrated mainly in the bottom and right part of the map.

It is possible now to examine the areas of overlap between the constructions, and identify their unique regions. Overall, the semantic region of *laten* is larger than that of *doen*. The areas of little or no overlap between the constructions are populated mainly by the exemplars of *laten*, whereas *doen* does not have a clear-cut conceptual area of its own. Thus, *laten* can be regarded as a hypernym of *doen*, rather than its synonym. The top left area does not contain any exemplars of *doen*, at least in this random sample. It is the zone with highly frequent coreferential Causers and Affectees. In these contexts the direction of the energy flow is usually reversed, going from the Causee to the Causer (Affectee). If the Causee is expressed, it is frequently marked with the preposition *door*, which usually marks agents, as in (4):

(4) *Je moet je niet door cijfers laten intimideren.*
you must you not by numbers let intimidate
"You shouldn't let yourself be intimidated by numbers." (*nl.beurs*)

The Causer is thus minimally involved in producing the outcome. Its role is reduced to that of the responsible entity (cf. Loewenthal 2003), which can resist or yield to the Causee's influence.

Another region where *laten* dominates is the area at the top. One of the very few exemplars with *doen* found here contains a Causee marked with *aan*, which is highly untypical of the construction. The choice of *doen* instead of *laten* construes the Causee as an incomplete agent, no longer able to control his or her actions. This inability is exploited by the Causer:

(5) *Een brief doen tekenen aan een mens van 90 jaar is*

a letter do sign to a person of 90 years is
niet moeilijk.
not difficult
“It is not difficult to make a person of 90 years old sign a letter.” (*be.finance*)

The other exemplars of *doen* in this region contain the idiomatic quasi-figurative expression *van zich doen spreken* “make speak about oneself, make one's mark”, which is in fact closely related to affective causation. The *laten*-exemplars predominant in this area contain Causees that intentionally bring about a change in another entity. In contrast with the previous example, the Causees act according to their will and interests, as in (6), which exemplifies the service frame:

(6) *Dit document kunt u bekomen op uw postkantoor*
this document can you receive on your postoffice
of laten opsturen door de Internationale dienst.
or let send by the International service
“You can receive this document at your post office or have it sent by the International service.” (*be.finance*)

Doen is also comparatively infrequent in the semantic area of providing information and showing something (a densely populated region on the left). The exemplars with *doen* which occur here contain *vermoeden* “suppose”, *geloven* “believe”, *vrezen* “fear” and some other Effected Predicates. Whereas the semantics of giving information involves the Causee as a relatively unaffected recipient or even a beneficiary, the causation patterns denoted by *doen* contain the Causees who are more cognitively and/or emotionally affected, in some cases even negatively. Compare (7) and (8), both with verbs of perception. In (7), the Causee might enjoy the caused event, and be willing to undergo it, whereas in (8) the Causees are affected by the causation in an extremely negative way.

(7) *Ik liet een vriendin een song horen.*
I let a friend a song hear
"I let a friend hear a song." (*De Standaard*, Oct. 2001)

(8) *Hij prijst de Palestijnse zelfmoordaanslagen en*
he praises the Palestinian suicide-attacks and
spreekt van 'lichamen die worden opgeblazen
speaks of bodies that are blown-up
en de Israëli's de dood doen proeven'.
and the Israeli the death do taste
"He praises the Palestinian suicide attacks and speaks about 'bodies that are blown up to make the Israeli taste the death'."
(*De Volkskrant*, March 2002)

The dense cluster with providing information is next to a relatively large constellation of *doen*-exemplars with the predominant *denken aan.* Note that mental associations are also a kind of perception (cf. Verhagen and Kemmer 1997: 73). In addition, they are closely related to physical perception: an object can remind one of something because it looks, sounds, smells, etc. in a particular way. However, in the case of *doen denken aan* the Causer plays the role of a stimulus, and the Causee is a cognizer who performs the effected mental event, regardless of his or her will. The causation is unavoidable, like in the behaviourist stimulus – response chain. This kind of perception is different from the processes encoded by *laten zien* or *horen*, which construe the act of perception as transfer of information from the source (the Causer) to the recipient (the Causee). The latter is not only unaffected, but also may be actively looking or asking for the information (e.g. *Laat me iets weten!* “Let me know something!”).

In the bottom central part of the map is a group of observations with both *doen* and *laten*, which contain metaphorical and quasi-metaphorical expressions with the physical source domain and mental or social target, such as *een belletje doen rinkelen* “ring a bell”, *de druppel die de emmer deed overlopen* “the drop that made the cup (lit. bucket) run over”, *iemand doen kokhalzen* “make someone sick (lit. make someone retch)”, *zijn oog laten vallen op iets* “spot something (lit. let one's eye fall on something)”, *het verdriet laten slijten* “let the grief wear off”, etc. The difference between the images underlying the events expressed with *doen* and *laten* concerns the actual cause of the event, or the main source of energy. In the case of the metaphors with *doen*, it is the Causer (an event or stimulus), whereas in the case of *laten* the actual cause is elsewhere: gravity, the natural course of events, etc. The role of the Causer is then reduced to that of the responsible entity.

The bottom right area contains exemplars with mostly physical caused events. The conceptual difference between the *doen*- and *laten*-exemplars found in this area involves again the primary source of energy, or cause: the Causer (*doen*), or another force or natural process (*laten*), e.g. gravity, mechanic energy, etc. (See Chapter 5, Section 5.3 for more examples). Compare (9) and (10):

(9) *In het stuk (...) All'aure in una lontananza (1977), worden de*
in the piece All'aure in una lontananza (1977) are the
oren door haast onhoorbare fluittonen minutenlang
ears by almost inaudible flute-tones for-minutes
uitgenodigd om zich wijd te openen (...),
invited in-order themselves wide to open
totdat een plotselinge schrille toon het trommelvlies dicht doet

until a sudden shrill tone the eardrum shut does

slaan.

become

“In the piece All'aure in una lontananza (1977) the ears are invited to open wide by almost inaudible flute tones, which sound for minutes, until a sudden shrill tone causes the eardrum to shut down.”

(*De Volkskrant*, Feb. 2002)

(10) (…) *dan laten ze een stuk van dien*d boot zinken*.

then let they a piece of that boat sink

“(...) then they let a piece of that boat sink.” (*CGN*, fv400729)

The next relatively large constellation of *doen* is found in the region of social caused events and abstract Causees in the right part of the map. All *doen*-Causers in this area are the actual causes of what happens, e.g. (11), whereas most exemplars of *laten* located in the neighbourhood describe the situations where the effect occurs due to the Causer's neglect or deliberate non-interference. The change may occur after the Causer stops the impingement, as in (12).

(11) *Verklaar mij anders eens wat het nut zou zijn*

explain me otherwise once what theuse would be

van de heksenjacht die de traditionele distributie in

of the witch-hunt that the traditional distribution in

kwaad daglicht stelt en zodoende de verkoop doet

bad daylight puts and thus the sales does

dalen.

go-down

“Explain to me otherwise what would be the point of the witch-hunt that puts the traditional distribution in bad light and by doing so causes the sales to go down." (*be.music*)

(12) *Door de handel worden de koersen gedurende een paar*
through the trade are the prices during a couple
jaren omhoog gemanipuleerd en daarna laten de
years up manipulated and then let the
effecternhandelaren de koersen zakken om
stockbrokers the rates fall in-order
vervolgens weer de koersen ophoog te zetten.
later again the prices up to set
"By trading, the prices are artificially pushed up during a couple of years and then the stockbrokers let the prices go down to raise them again later." (*nl.financieel.beurs*)

Finally, the top right segment of the map is a region with mostly intransitive social and physical caused events performed by human Causees. The difference between the few *doen*-exemplars and the observations with *laten* is that the former involve forcing the Causee to carry out the desired action, as in (13), whereas the latter tend to refer to the situations when the Causee is expected to perform the action willingly, as in (14).

(13) *Waar haalt enig politiek vertegenwoordiger nog enig recht*
where gets any political representative still any right
om welke belg ook om het even wat op te leggen of
in-order which Belgian too no matter what up to lay or
dwingend te doen opvolgen?
forcefully to do obey?
"Where does any political representative get any right to impose anything on any Belgian or force him/her to obey?"(*be.politics*)

(14) *(...) ik heb met deze Leekens-mannetje sindsdien geen*
I have with this Leekens-man since-then no
illusies meer over zijn totaal gebrek aan psychologie
illusions more about his total lack to psychology
om rode duivels van verschillende landen te kunnen
in-order red devils of different countries to be-able
motiveren en samen te kunnen laten spelen.
motivate and together to be-able let play
“I don't have any illusions about this Leekens guy since then about his total lack of psychology to be able to motivate the Red Devils from different countries and have them play together.” (*be.sport.football*)

To summarize, the non-overlapping area with the reflexive constructions contains examples of the maximal autonomy of the Causee, and the minimal Causer's impact on the other participants. The difference between the constructions in the overlapping conceptual areas seems to be in the construal of causation as direct or indirect. The typical *doen*-Causer is the actual cause of the change, or the ultimate source of energy, and the typical *doen*-Causee has some properties of a patient (sometimes even of a victim). In the case of *laten* the Causer exploits voluntarily or involuntarily other energy sources, including natural forces and processes, machinery, the Causee's will, etc., and the Causee is frequently volitional and agent-like.

In general, the realization of the direct/indirect causation distinction and the specific roles played by the Causer and the Causee depend on the specific semantic area. For instance, in the mental causation area the main participants of the *laten*-construction are construed as a source of information and an addressee, and the Causer and the Causee of *doen* are frequently a mental stimulus and an experiencer. In the area of physical caused events, the distinction between indirectness, as opposed to directness, is frequently associated with the use of other sources of energy to bring about the caused event. As for social caused events, indirectness can involve neglect or non-interference in the current affairs. In the case of human Causees, indirect causation is based on the actions of willing motivated Causees, whereas direct causation often overrides a human Causee's desires and interests and even causes harm.

6.3. Distinctive features of *doen* and *laten*

6.3.1. Exploratory analysis

The next step was to compare the distributional profiles of *doen* and *laten* in order to find the features with the highest cue validity. I apply the same method of comparing the average proportions of each semantic feature in the samples of *doen* and *laten*, as the one that was used for comparing the clusters in Chapters 4 and 5. The snake plot in Figure 6.2 (an idea from Gries and Otani 2010) shows the 15 most distinctive features of *doen* and *laten*.

In the bottom right part one can find the most distinctive features of the construction with *doen.* Some of them echo the direct/indirect dimension, which was discussed above. These are intransitive Effected Predicates (*EPTrans.Intr*) and explicit NP Causees (*CeSynt.NP*), undergoing change unintentionally (*CeRole.Change*, *CeIntent.No*). The distinctive features of *laten* can be found in the top left part. Those related to indirectness are transitivity of the Effected Predicate (*EPTrans.Tr*) and corresponding lack of prepositional government (*EPPrep.None*), as well as intentional (*CeIntent.Yes*), implicit (*CeSynt.Impl*) and semantically less determinate human Causees (*CeSem.HumUndef*) that cause a change in another entity (*CeRole.Cause*). The higher frequency of pronouns as

Affectees (*AffPOS.Pron*) is explained by the high frequency of the reflexive pronouns in the coreferential contexts (*Coref.CrAff*).

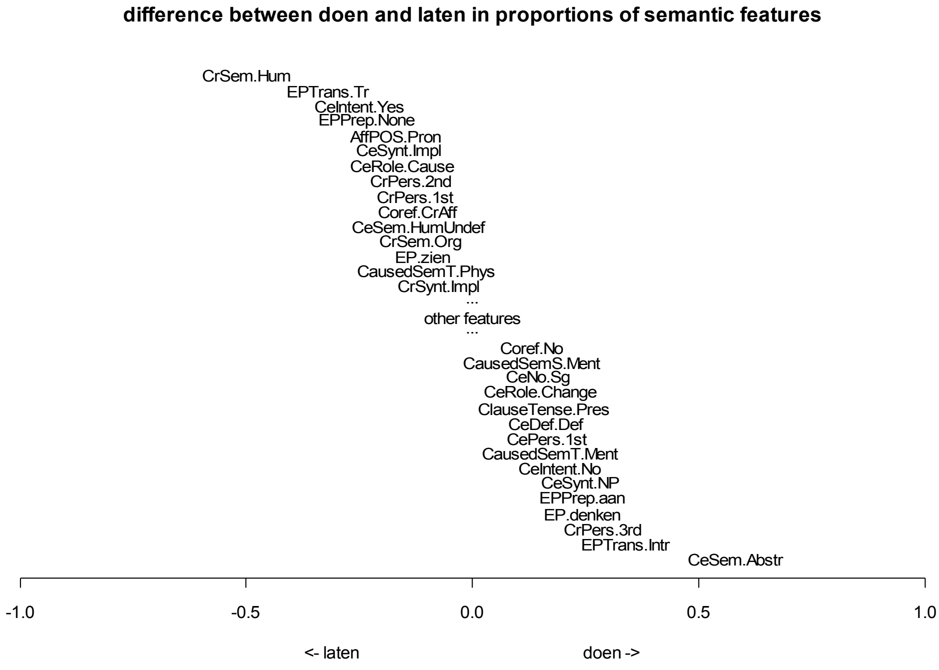


Figure 6.2. Top 15 most distinctive features of *doen* and *laten*, according to the proportions of the features in the constructional exemplars.

However, there is another important distinction that has not been captured before. The most distinctive feature of *laten* is the human Causer (*CrSem.Hum*), whereas *doen* is usually accompanied by abstract Causers (*CrSem.Abstr*), hence the 1st and 2nd person Causers for *laten* vs. 3rd person Causers for *doen*. According to the previous studies of direct and indirect causation, human beings normally act upon other human beings indirectly (Verhagen and Kemmer 1997). It also seems that humans are more capable of affecting things indirectly, too, i.e. by using tools, natural forces, and predicting the outcome of their own actions and other events. In fact, one could regard indirect causation as a distinguishing evolutionary feature of human beings. Moreover, abstract Causers are perfectly fit to be the *causes* of the caused event – in other words, to represent the causing event or state itself, as well as the stimuli that trigger mental associations. Recall that abstract social causation and affective causation are the most entrenched senses of *doen*. Human Causers, on the other hand, are better fit to bear moral and social responsibility for causation because they are supposed to be in control of the world and themselves, and, as it has been demonstrated, the role of the responsible entity is crucial for the Causer in the *laten*-construction, especially in the exclusively *laten*-area with the coreferential Causers and Affectees.

Another important difference is that *doen* is strongly associated with mental caused events (*CausedSemT.Ment* and *CausedSemS.Ment*). However, the previous analyses make one suspect that this preference may be due to the large cluster of *doen denken aan*. The high relative frequency of this expression explains the distinctiveness of these features. The 1st Person singular pronouns as the Causees (*CePers.1st*, *CeNo.Sg*) and the present tense in the clause (*ClauseTense.Pres*) are signs of the speaker's involvement and subjectivity commonly associated with this sense. The top features of *laten*, though, include another mental predicate *zien* “see” – due to the large cluster of providing information and showing.

6.3.2. Confirmatory multivariable analysis

In this subsection, I test a) whether the results of the feature comparison can be extrapolated to the entire population of the constructions at an acceptable level of statistical significance, and b) if the features that were found to be important in the individual tests will behave in the same way if the other features are taken into account. To answer these questions, I carried out several confirmatory multivariable analyses.

First, I fitted a logistic regression model with *doen* or *laten* as the response variable and the following predictors:

- *CrSem*, or the semantic class of the Causer: human individuals, organizations, undefined human participants (individual persons or organizations), material objects and abstract entities. The remaining values were conflated in these five in order to avoid data sparseness;

- *CeSem*, or the semantic class of the Causee: the same values as for the Causer;
- *CeSynt*, or the syntactic expression of the Causee: explicit zero-marked NP, explicit prepositionally marked NP, or implicit;
- *CeIntent*, which specifies whether the Causee acts intentionally (yes, no, or undefined);
- *CeRole*, or the role of the Causee with regard to the caused event specified by the Effected Predicate: the Causee causes a change expressed by the verb, undergoes a change, or there is no change at all;
- *Coref*: coreferentiality of the Causer with the Causee, with the Affectee, or no coreferentiality at all;
- *EPTrans*: a transitive (including ditransitives) or intransitive (including indirect object constructions) Effected Predicate;
- *CausedSemT*, or the target semantic domain of the caused event: physical, social or mental.

These variables represent two main dimensions of semasiological and onomasiological variation of *doen* and *laten*. The distinction between mental and non-mental caused events is represented by *CausedSemT* and *CeSem* (mental caused events always involve individual human Causees). One can expect *doen* to be favoured by mental caused events and individual human Causees, according to Figure 6.2. The second dimension, arguably the more important one, is directness and indirectness of causation. One can expect the features associated with directness (abstract Causers, intransitive Effected Predicates, Causees undergoing a change unintentionally, etc.) to increase the probability of *doen*, and the ones interpreted as signs of indirectness to boost the chances of *laten*. Of course, more variables could be added to the model. However, the other candidates either contained too many missing values (e.g. the variables describing the Affectee), or were strongly associated with the above-mentioned variables (e.g. prepositional complements are typical of intransitive verbs; 1st and 2nd person pronouns normally refer to human beings), which would make the estimates unreliable.

I also controlled for the lectal variation by including the Country and Register in the model. According to the results of the previous studies (see Section 2.3 of Chapter 2), Belgian Dutch and the more formal registers (the newspapers, and to a lesser degree the Usenet) were expected to increase the probability of *doen*. These variables were included in order to neutralize a possible confounding effect of the varieties on the use of the constructions.

The previous analyses also revealed several highly frequent collocational patterns in the use of the causatives (e.g. *doen denken aan*), which can trigger exemplar effects in categorization. In other words, the speaker's choice between the auxiliaries may be affected by these entrenched set expressions. As we saw in the previous chapters, the frequencies of these patterns differ in the lectal subsamples because the specific collocations are closely related to the communicative goals and referential situations associated with every lect (see also Levshina et al. Submitted). As a consequence, the generalizations derived from the local patterns may differ if we use another corpus. Therefore, one should control for the low-level schemata when testing the generalizations.

A possible solution is to use mixed-effect models, which allow a statistician to separate these sources of variation. Mixed-effect models contain both fixed effects, like the variables listed above, and random effects – most commonly, variables that reflect interrelatedness of the observations, such as the subcorpora from which different sets of observations come, specific linguistic stimuli presented in an experiment several times, or the same subjects that perform several tasks (see Baayen 2008 for linguistic examples). In our case, mixed models will filter out the lexical effects to check whether the generalizations will hold if we use other data (cf. Bresnan et al. 2007, who do the same with specific verb senses to test the global factors that influence syntactic choices).[1]

The random effects in my mixed-effect model were the Effected Predicates. There were 1058 types in total, with the token frequency from 1 to 549 (*zien* "see"). I also treated *denken* + *aan* apart from *denken* because *denken aan* as the Effected Predicate has a meaning "remember, think of", which stands apart from the other uses of the predicate. I used the entire data set for the test, omitting the observations with missing values. This resulted in the total of 5548 exemplars. The predictive power of the model was excellent, with C = 0.986, Somers' D_{xy} = 0.973, and the pseudo-R^2 = 0.766.[2]

The estimates of variables in the model without interactions are shown in Table 6.1. They represent the log odd ratios of *doen* vs. *laten.* The positive estimates indicate the features that boost the probability of *doen* in the given context, whereas the negative estimates show that the feature increases the chances of *laten*. The table does not show the reference level of the variables, i.e. the value with which all the other values are compared. These were the following values: abstract Causers, abstract Causees, explicit zero-marked NP as the Causee, Causees acting unintentionally, no observable change, no coreferentiality between the Causer and the other participants, intransitive Effected Predicates, mental caused events, the Netherlandic variety and newspaper register. The estimates for these levels can be considered equal to 0. The smaller the *p*-

[1] Of course, I do not assume that the verbs occur in the Effected Predicate slot randomly, but their relative frequencies may depend on the sample.

[2] C: the area under the ROC curve, or probability of concordance between predicted and observed responses, usually in the range from 0.5 (random prediction) and 1 (perfect prediction). Somers' D_{xy}: rank correlation between predicted probabilities and observed responses in the range from 0 to 1. Pseudo R^2 : the squared correlation between the observed outcome and predicted values) in the range from 0 to 1.

value, the more confident we can be that the estimate differs from 0.

Feature	Estimate	*p*-value
(Intercept)	1.602	< 0.001
CrSem = Human	- 3.759	< 0.001
CrSem = Human Undefined	- 3.738	< 0.001
CrSem = Material Object	- 0.545	0.25
CrSem = Organization	- 4.017	< 0.001
CeSem = Human	0.612	0.051
CeSem = Human Undefined	0.997	0.073
CeSem = Material Object	0.038	0.913
CeSem = Organization	0.261	0.54
CeSynt = Implicit	- 0.326	0.042
CeSynt = Prepositional Phrase	0.318	0.585
CeIntent = Yes	- 0.777	0.014
CeIntent = Undefined	0.154	0.718
Coreferentiality of Causer and Causee	- 1.545	0.056
Coreferentiality of Causer and Affectee	- 2.871	< 0.001
EPTrans = Transitive	- 1.772	< 0.001
CeRole = Cause	0.356	0.452
CeRole = Change	0.528	0.034
CausedSemT = Physical	- 1.082	0.001
CausedSemT = Social	- 0.596	0.044
Country = Belgium	0.521	0.002
Register = Conversations	- 0.959	0.002
Register = Usenet	- 0.38	0.038

Table 6.1. Estimates in a mixed-effect logistic regression model (only main effects).

The estimates show that most expectations based on the previous analyses are borne out. In addition, some new details come to light. The features related to the direct and indirect causation distinction behave in the following way:

- *CrSem*. The semantic class of the Causer is the strongest predictor, according to the the magnitude of the estimate. All human Causers, including undefined humans and organizations, particularly disfavour *doen*. Abstract Causers (the reference level) are then the best Causers for *doen.* Material Causers do not differ in their preferences from abstract Causers at a statistically significant level (here and in what follows $\alpha = 0.05$);
- *CeSynt*. Overall, implicit Causees prefer *laten* more than explicit zero-marked NPs (the reference level), although the difference has borderline significance. Prepositionally marked Causees do not show a significant effect, when the other factors are controlled for;
- *CeIntent*. As expected, intentionally acting Causees boost the probability of *laten*, as opposed to unintentionally acting Causees. Undefined intentionality does not differ significantly from lack of intentionality;
- *CeRole*. Contrary to the expectations, when all other factors are controlled for, the contexts with the Causees causing a change are not significantly different from lack of any change (the reference level) with regard to the choice between *doen* and *laten*. However, if the Causee undergoes a change itself, there is some increase in the probability of *doen*, as expected;
- *Coref.* Coreferentiality of the Causer and Affectee, which reverses the energy flow back to the Causer, seems to be a very strong predictor of *laten*. Recall that in the MDS solution (Figure 6.1) the region with this type of coreferentiality is populated only by exemplars of *laten*. In addition, if the Causer and Causee are coreferential (*Ik liet me gaan* "I let myself go"), this also increases the chances of *laten* in comparison with the reference level (no coreferentiality), although with borderline statistical significance;

- *EPTrans*. Transitivity of the Effected Predicate is yet another powerful pro-*laten* factor, whereas intransitive verbs are more tolerant of *doen*.

The other dimension, mental vs. non-mental caused events, highlights the prominence of affective causation for *doen*, albeit not very strongly. Individual human Causees boost the probability of *doen*, but their estimate has a borderline significance. Physical caused events (variable *CausedSemT*) and to a lesser degree social caused events disfavour *doen* more than mental caused events do.

The lectal effects echo the tendencies found on the basis of the *doen*/*laten* ratios: *doen* has higher chances to occur in Belgium and in more formal registers. The analysis shows a decline of probability of *doen* from more public, formal and prepared speech to more informal, private and spontaneous communication. More on this will follow in the next section.

Let us now have a look at the lexical component of variation. The variance of the random effects was substantial: 3.12, with the standard deviation 1.77. Table 6.2 displays the top 5 greatest adjustments that increase the probability of *doen*, and those that boost *laten*. Every adjustment is the coefficient that is added to the model if the observation contains the given verb.

These verbs are Effected Predicates that are frequently found in untypical contexts with *doen* or *laten*. The verbs *zien*, *horen* and *weten* denote mental states, which are the least typical of *laten*, according to the fixed effect estimates. On the other hand, *denken aan* freely combines with human Causers and frequently has implicit Causees (see Section 4.3 in Chapter 4) – the features that boost the probability of *laten*. The verb *voorkomen* "appear" occurs frequently with human Causers, as in (15), although the latter are not typical of *doen*. The verb *wachten* "wait" is intransitive, and frequently occurs with abstract Causers in the expression *op zich laten wachten* “be long time coming (lit. let wait for oneself)”, as in (16). The adjustments compensate for such deviations from the typical distinctive features of *doen* or *laten*. Most of these differences are due to the relatively autonomous status of the fixed expressions, which exhibit their own idiosyncratic properties (cf. Bybee 1985). Mixed models keep these idiosyncrasies apart from the 'normal' tendencies, preventing them from influencing the estimates of the fixed effects. The method is thus helpful in the identification of semantically autonomous low-level schemata.

doen	***laten***
vòòrkomen “appear, look (as if)” 5.37	*zien* “see” - 3.43
denken aan “think of” 4.746	*liggen* “lie” - 3.10
vergeten “forget” 4.48	*horen* “hear” - 2.89
verzorgen “provide, take care” 4.30	*wachten* “wait” - 2.89
vinden “find” 4.04	*weten* “know” - 2.52

Table 6.2. Top 5 pro-*doen* and pro-*laten* adjustments for Effected Predicates treated as random effects.

(15) *Zo deed Cziommer het wel voorkomen voor*
so did Cziommer it well appear in-front-of
de camera's.
the cameras
“Cziommer made it look like this in front of the cameras.” (*nl.sport.voetbal*)

(16) *De tegenactie liet bijna een maand op zich*
the counteraction let almost one month on itself
wachten.

wait

"The counteraction was delayed for almost a month." (*De Morgen,* Oct. 2001)

Although the model supports the previous findings, it is necessary to check possible interactions between the factors. An interaction is observed if the joint effect of two or more variables is different from the sum of their individual effects. Modelling interactions between the factors in the model ran into the problem of high-order interactions and, consequently, cells with zeros (even after the levels of some factors were conflated), which made the model unreliable. To solve this problem, I used a non-parametric approach of conditional inference trees (see Tagliamonte and Baayen, Submitted). This method is based on binary recursive partitioning and works as follows: 1) the algorithm tests if any of the variables are associated with the given response variable (*doen* or *laten*), and chooses the variable that has the strongest association with the response by measuring the *p*-value of this association; 2) the algorithm makes a binary split in this variable; 3) the first two steps are repeated until there are no variables that are associated with the outcome at the pre-defined level of statistical significance. The result of this process can be visualized as a dendrogram. An important difference between this method and the more traditional regression and classification trees (e.g. Heylen 2005) is that the conditional inference trees are more robust with regard to the different number of factor levels. The method is realized as the `ctree` algorithm in the `party` package in R (Hothorn et al. 2006).

Figure 6.3 presents the tree with all possible splits, significant at $\alpha = 0.05$. All variables from the logistic regression model were tested, plus the twenty-five most frequent Effected Predicates, which included *denken aan*, *zien*, *weten*, etc. (the rest were coded as "Others"). The ovals contain the names of the variables selected for the best split, as well as the *p*-values.

The levels of the variables are specified on the “branches”. The bars at the bottom ('leaves') show the proportions of *doen* (black) and *laten* (gray). The numbers above the bars show the number of observations in each end node. The statistics demonstrate that the tree separates *doen* from *laten* very successfully ($C = 0.95$, $D_{xy} = 0.91$).

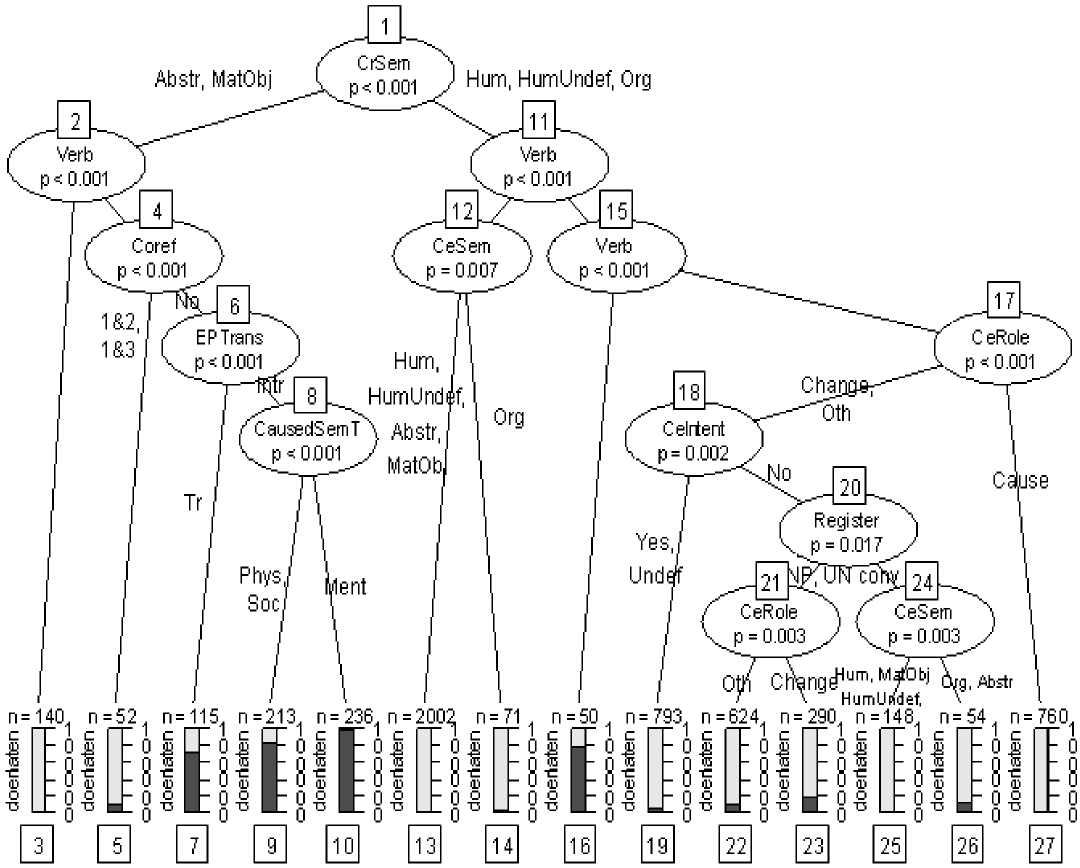


Figure 6.3. Conditional inference tree with all factors based on a Monte-Carlo test with 9999 replications. Splits were made if p < 0.05. Minimal number of observations in terminal nodes equaled 50. The numbers in squares are node IDs.

From the figure one can imagine the complexity of the relationships between the predictors. Some of them participate in a split several times. The most important predictor (the top split) is, again, the semantic class of the Causer. The left part of the graph (abstract and material Causers) contains some bars with predominant *doen*, whereas the right part (all kinds of human Causers) on average favours *laten*. In fact, there are only a few bars with a higher proportion of *doen.* Most of them are located on the branch with inanimate Causers. This fully corroborates the previous findings. Next, the individual verbs come into play (not shown due to lack of space), followed by some structural, semantic and lectal factors. In some cases the lexical factors can override the power of the Causer's semantics. For instance, one can find one bar with predominant *doen* (Node 16) on the branch with human Causers: it is restricted to the Effected Predicates *denken aan* "remember, think of", *geloven* "believe", *vermoeden* "suppose" and *stijgen* "rise".

Nodes 4 and 17 are those that contain the majority of verbs ('Other') after most of the high-frequency verbs were 'peeled off' in the previous splits. This is where the general semantic and other factors come into play. The leaves that are associated with coreferentiality (Node 5), transitivity (Node 7), non-mental caused events (Node 9), active role of the Causee (Node 27), intentionality of the Causee's actions (Node 19) and the conversational register (Node 24) all disfavour *doen* more than their sister nodes do, but this difference manifests itself in the presence of many other features. Therefore, there is evidence of complex multiway interactions between the variables. Interestingly, the semantics of the Causee has some effect in Nodes 24 and 12, but in both cases it forms the last split on the branch.

The graph demonstrates that lexical fixation is very influential. When the collocational information is not available (for low-frequency verbs and hapax legomena), the more abstract semantic and pragmatic factors seem to come into play. This may reflect the speaker's actual categorization process, although the hypothesis needs further testing.

6.4. Lectal differences in the division of semantic labour between *doen* and *laten*

As in the previous chapter, the analysis of lectal variation begins with an overview of the quantitative differences in the use of *doen* and *laten* in the six lects. Figure 6.4 displays the *doen/laten* ratio in each lect. One can see that the largest proportion of *doen* is found in the Belgian newspaper data, whereas the Netherlandic spontaneous conversations yield the smallest proportion of *doen*. This corroborates the results of the logistic regression in the previous section, which showed that *doen* is more frequently preferred (or rather, less frequently avoided) in more formal communication and in Belgium.

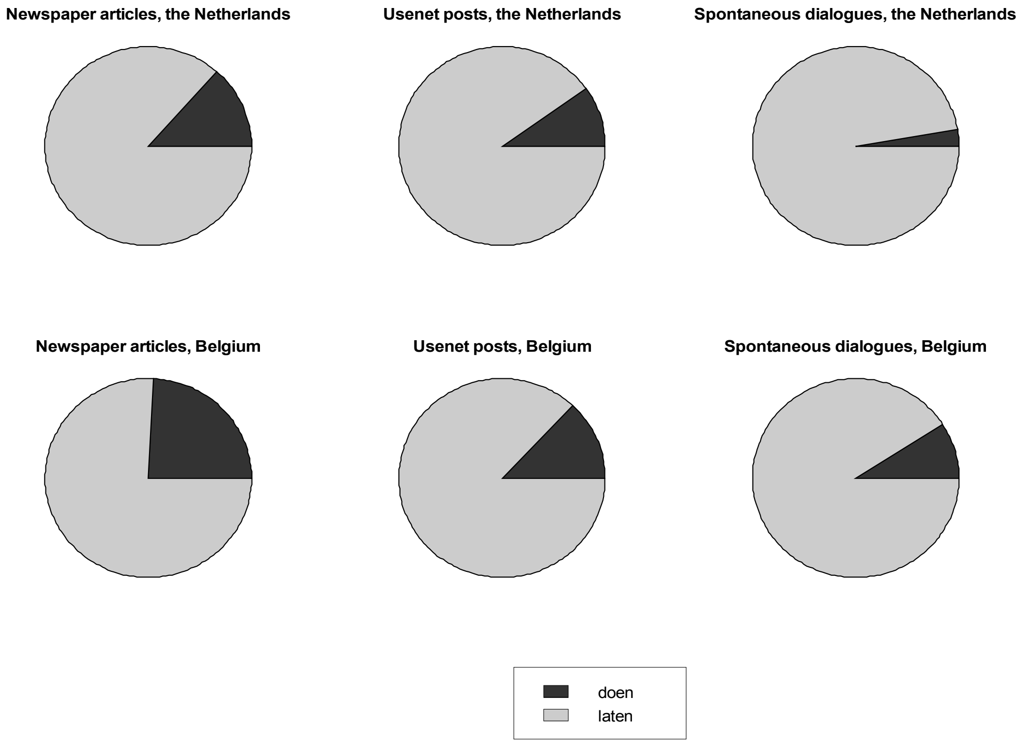


Figure 6.4. Relative frequencies of *doen* and *laten* in six lectal samples.

Interestingly, the proportions of *doen* in the pie charts resemble the individual frequencies of the construction in Figure 4.6. For *laten*, no such correlation is observed. It might be that the overall normalized frequency of *doen* in the corpus depends on how successfully the construction competes with *laten*, whereas there are no clear indications of that for *laten*. The overall normalized frequency of the latter seems to depend more on 'extralinguistic' factors, such as the referential situations conveyed by the speakers and their communicative goals.

On the basis of the results of the lectally enriched semasiological analyses in Chapters 4 and 5 one can make several predictions with regard to the division of labour between *doen* and *laten*. First, if the affective causation pattern with *doen denken aan* is more frequent in the Netherlandic and conversational data, then one could expect the corresponding lexical effects and mental caused events in general to be more distinctive of *doen* in these lects. One can also expect a stronger effect of transitivity in the Netherlandic newspapers, especially in comparison with the Belgian ones, and a weaker effect of the mental/non-mental distinction, because the Netherlandic newspaper sample contains a very high proportion of transitive verbs of perception and knowledge as the Effected Predicates of *laten* in comparison with the Belgian newspapers.

To test these expectations, I calculated the differences between the proportions of the features for *doen* and *laten* in all six lects. The top five distinctive features of *doen* and *laten* are shown in Table 6.3. As one can see, the expectations are borne out. First, transitivity (*EPTrans*) has higher ranks both in the top five of *doen* and that of *laten* in the Netherlandic newspapers than in any other lect, except for the Belgian Usenet (although the difference in proportions, which is not shown, is smaller in the latter), whereas the feature is not even listed in the top five of *laten* in the Belgian newspapers. Second, *denken aan* (*EP.doen* and *EPPrep.aan*) is more distinctive of *doen* in the Netherlandic registers than in the Belgian lects, and slightly more in the spoken data than in the written registers in both countries. The expression is very frequently used with the personal pronoun *mij* “me” in the Causee slot (*Het doet me denken aan X* “It reminds me of X”) in the Usenet and conversations. This is why the 1st person Causees are highly distinctive of the *doen*-construction in the Netherlandic Usenet and both conversational samples, but not in the newspapers. Finally, the predominance of physical caused events with *laten* in the conversations is also observed in the results (see *CausedSemT.Phys* and *CausedSemS.Phys*).

Rank of feature	**News NL**	**News BE**	**Usenet NL**	**Usenet BE**	**Convers. NL**	**Convers. BE**
doen 1	CrSem. Abstr	CrSem. Abstr	CrSem. Abstr	CrSem. Abstr	EPPrep.aan	CrSem. Abstr
doen 2	EPTrans. Intr	CeIntent. No	CePers.1st	EPTrans. Intr	EP.denken	EP.denken
doen 3	EP.denken EPPrep.aan	CeRole. Change	EP.denken	CrPers.3rd	CrPers.3rd	EPPrep.aan
doen 4		CausedSem T.Ment	EPTrans. Intr	CeSynt.NP	CausedSem T.Ment	CePers.1st
doen 5	Adv.None	EPTrans. Intr	EP.Prep.aan	ClauseTens e.Pres	CePers.1st	CausedSem T.Ment
…	…	…	…	…	…	…
laten 5	CeIntent. Yes	CausedSem T.Soc	EPPrep. None	CeSynt. Impl	CeIntent. Yes	CausedSem S.Phys
laten 4	Clause. Main	CrSem.Org	CeSynt. Impl	CrPers.2nd	EPTrans.Tr	CausedSem T.Phys
laten 3	EPPrep. None	CausedSem S.Soc	CeIntent. Yes	CeIntent. Yes	CausedSem T.Phys	EPTrans.Tr
laten 2	EPTrans.Tr	CeIntent. Yes	EPTrans.Tr	EPTrans.Tr	EPPrep. None	EPPrep. None
laten 1	CrSem. Hum	CrSem. Hum	CrSem. Hum	CrSem. Hum	CrSem. Hum	CrSem. Hum

Table 6.3. Distinctive features of *doen* in six lectal samples. Merged cells contain the features with the same rank.

Interestingly, the position of the construction in the main clause (*Clause.Main*) is distinctive of *laten* in the Netherlandic newspapers because *doen* is very frequently used in relative clauses of the type *X dat/die doet denken aan...* “which reminds of”, especially in the articles about music (see Section 4.3 of Chapter 4).

In addition, *laten* in the Netherlandic newspapers is very frequently accompanied by adverbial modifiers of time, especially in the contexts like *De minister liet gisteren weten dat...*"The minister said yesterday that...". As a result, the absence of adverbial modifiers (*Adv.None*) is a distinctive feature of *doen*.

One can conclude that the general division of conceptual labour between the two constructions is similar across the lects. The differences in the distribution of specific features can be explained by the frequently used lexical expressions, specific referential situations and communicative functions of the registers.

There are two popular confirmatory approaches to modelling the relationships between the semantic and lectal factors in an onomasiological analysis. First, one can fit different models for each lect and then compare the estimates of the linguistic factors in the different models (e.g. Grondelaers et al. 2002; Szmrecsanyi 2010). The other method is to incorporate both the cognitive and lectal factors in one model and test their interactions (e.g. Bresnan and Hey 2008; Speelman and Geeraerts 2009; Bresnan and Ford 2010). The presence of interactions is a sign of potential lectal difference in the division of conceptual labour between the constructions. However, as was mentioned above, modelling interactions in this case is problematic due to complex high-order interactions and, consequently, many zero cells in the data. Fitting separate models will face the problem of data sparseness, too, especially in the case of the spoken and Netherlandic data with low frequencies of the *doen*-response. This is why I used a relatively novel technique of random forests for each lectal subsample (see Tagliamonte and Baayen, Submitted). This technique allows for modelling the conditional importance of each predictor in the situations of data sparseness, high-order interactions, and highly correlated predictors. The conditional importance approach is a robust method of measuring the individual weight of variables that takes into account correlations between the variables (Strobl et al. 2008). A random forest is grown from a set of conditional inference trees, which were discussed above. Variable importance measures are calculated for every individual tree and then averaged over all trees in the forest.

Appendix 2 shows the random forests grown for each of the six lects. The classification power of the models was excellent again, with the minimum $C = 0.939$ and $D_{xy} = 0.879$ for the Belgian newspapers and the maximum of $C = 0.988$ and $D_{xy} = 0.976$ for the Netherlandic conversations. The number of trees in each forest was 1000.[1] The horizontal axis presents the variable importance index for each variable. The dashed line indicates the level of significance – traditionally, it is the absolute value of the smallest variable. In all the models presented, it is zero or very close to zero. All models show that the semantic class of the Causer plays by far the most important role. As for the other variables, there is variation in the relative importance. In the spoken data, both in Belgium and the Netherlands, the next important variable is the semantic domain of the caused event, which is followed by transitivity. This supports the observation made previously that the *doen*/*laten* distinction in spoken data is to a larger extent motivated by the opposition between the mental (*doen*) – physical caused events (*laten*) than in the written registers. The written registers in the Netherlands have an outspoken transitivity effect. This might be explained by the high frequency of *laten weten*, *laten zien* and some other transitive fixed expressions discussed earlier (this is also supported by the results in Levshina et al, Submitted). The Belgian Usenet follows this pattern to some extent, but the Belgian newspapers idiosyncratically boost the relative effect of the role of the Causee in the event denoted by the Effected Predicate, although transitivity and intentionality are significant, too. This effect is due to the high number of *doen*-exemplars denoting quantitative change of the Causee in the Belgian newspapers (see Section 4.3 in Chapter 4). Checking the conditional

[1] The tuning parameter *mtry* was set at 3 (see Strobl et al. 2008).

inference trees for every lect (not shown) and examining the splits and their marginal effects demonstrates that the effects of the variables are similar in all lects. The trees differ mainly in the number of branches due to the unequal numbers of exemplars in the samples. Thus, we can consider the results of the exploratory analyses confirmed.

6.5. Summary

This chapter focused on the distinctive features and senses of *doen* and *laten*. The main findings are as follows.

- A simple measure of cue validity of the exemplars of *doen* and *laten* with regard to their categories has shown that the most distinctive exemplars of *doen* are the ones that belong to the *doen denken aan* pattern. The lowest cue validity scores for *doen* and the highest ones for *laten* are obtained for the exemplars that denote inducive causation (the service frame). The least distinctive patterns of *laten* are those with abstract Causers and intransitive Effected Predicates. The fact that the configurations with the lowest cue validity scores of *doen* are very similar to the ones with the highest cue validity of *laten*, but that the reverse does not hold, might be due to the exemplar effects of *doen denken aan*.
- The most important distinctive dimension is that of direct and indirect causation, associated primarily with the role of the Causer and the Causee in causation. As the previous studies suggested, *doen* tends to express more direct causation, whereas *laten* is the indirect causation auxiliary. These tendencies are probabilistic rather than categorical due to a significant overlap of the conceptual areas associated with the two constructions, although there is an area where *laten* dominates absolutely (transitive constructions with

coreferential Causers and Affectees). This implies that the relationships between *doen* and *laten* in general are those of hypo- and hypernymy. I have also identified some local differences in the construal in the areas of overlap, which manifest themselves in specific roles of the main participants, such as the source of information and the addressee, or the stimulus and the experiencer.

- Another important distinction is that between mental and non-mental caused events. *Doen* tends to express more mental caused events than *laten* (affective causation), although this is to a large extent due to the high entrenchment of *doen denken aan* "remind of" and other fixed collocations. In the confirmatory mixed-effect regression model this dimension had only borderline significance.
- The lexical effects play a very important role in the choice between *doen* and *laten*. This is obvious from the mixed effect model and conditional inference tree. Some Effected Predicates can override the general 'prototypical' tendencies in the division of labour between the constructions. There is therefore evidence of exemplar effects in categorization of causative events.
- I have also found some lectal differences in the division of labour between the two constructions in a series of exploratory and confirmatory analyses with the help of simple proportions and random forests. To a large extent, they can be explained by the differences in the referential situations and communicative functions performed by the lects, and the entrenched collocational patterns which were identified in the previous chapters.

Chapter 7. Discussion and perspectives

In this final chapter I discuss the main findings of the previous chapters from the point of view of the central theoretical and methodological issues which were raised in the Introduction: the relationships between the intra- and intercategorial perspectives (Section 7.1), the interaction of conceptual and lectal factors of language variation and change (Section 7.2) and, finally, the interpretation of the results from a broader theoretical perspective (Section 7.3).

7.1. All sorts of salience

The results of the previous analyses revealed intra- and intercategorially salient semantic exemplars and features. Whereas the overall semantic structure of the categories seemed to be quite stable in the semasiological and onomasiological analyses (see the separate and common exemplar maps in the previous chapters), the salience scores of the same exemplars differed in some cases. What are the relationships between these different manifestations of salience?

According to the analyses, the intracategorial prototype of *doen*, operationalized as a combination of the most frequent features shared by most exemplars, was very close to affective causation with abstract Causers, mental caused events and human Causees, and had a strong bias towards *doen denken aan*. This type of meaning also corresponded to the most densely populated cluster on the semantic map. Note, however, that the centre of the exemplar map was sparsely populated, which suggests that the single generalized prototype of the category is quite schematic. The least intracategorially typical sense of *doen* was inducive interpersonal causation with intentional Causees and transitive Effected Predicates, located in a sparsely populated region of the semantic space. A very similar picture was observed when I performed the intercategorial analysis. The exemplars with *doen denken aan* were found to be the ones with the highest cue validity, and the inducive causation exemplars were again found the least distinctive of *doen*.

Thus, we can see that these salience phenomena are correlated in the case of *doen*. To test this conclusion, I performed a series of correlation tests between the four measurements of salience described in Table 3.5, Chapter 3. For the exemplars, the correlation (Pearson's product-moment coefficient) between the intracategorial family resemblance and intercategorial cue validity scores was 0.83, $p < 0.001$. I also compared the intracategorial weight and intercategorial distinctiveness of the features. The correlation was weaker, but still positive and statistically significant ($r = 0.44$, $p < 0.001$). These results are in line with the previous experimental and corpus-based studies of lexical categories, which showed that intra- and intercategorial salience of category members correlate positively (Rosch and Mervis 1975; Geeraerts et al. 1994). The weaker correlation between the feature-related salience indicators is due to the very low frequencies of some features in both constructions.

With *laten* the situation was more complicated. The intracategorial prototype (the configuration of the most popular semantic and other features) was even more schematic than that of *doen*, and included fewer specific features – most importantly, human Causers. There was also no specific non-prototypical sense: the exemplars with low family resemblance scores were very diverse. The most densely populated areas on the semantic map of *laten* corresponded to the senses of providing information and showing, which involved mainly the Effected Predicates *weten* “know”, *zien* “see” and *horen* “hear”, and the area of social caused events with the Effected Predicates *liggen* “lie”, *vallen* “fall” and a few others. The least populated areas were on the periphery around the cloud of exemplars and in the centre. From the intercategorial perspective, the most distinctive exemplars had transitive physical caused events and human undefined volitional Causees. Very commonly these exemplars were associated with the service frame. Many of the observations in the top fifty most distinctive exemplars also had coreferential Causers and Affectees. This suggests that the most distinctive senses of *laten* involve the affected Causer – a beneficiary or a victim of the Causee's actions, construed as the Causer because he or she is the ultimate responsible entity (and also a facilitator or an obstructor). The measurements of the distinctive features as differences in proportions (see Figure 6.2 in Chapter 6) yielded the features that were also found in the most distinctive semantic exemplars mentioned above, but there were no actual observations with *laten* that could fit the least distinctive features – those associated with the *doen denken aan* schema – due to the strong exemplar effects of the latter, which ensures the purity of the cluster from the contrasting category.

Thus, we can state that the most typical and the most distinctive features and exemplars of *laten* do not coincide. To test this, I performed the correlation tests for *laten*. For the exemplars, the correlation between the intracategorial family resemblance and intercategorial cue validity was negative and weak: r = -0.245, p < 0.001. As far as the features are concerned, there seemed to be no correlation at all: r = -0.059, p = 0.13.

In addition, I repeated the same tests for every lect individually. The correlation coefficients sometimes differed significantly, but the tendency remained very much the same. In every lect, the correlation coefficients for *doen* were much higher than those for *laten*. Normally, the intra- and intercategorial salience indicators yielded high or moderate positive correlations in the case of *doen*, and displayed lack of correlation, or even negative values for *laten*.

How could one explain this lack of correlation? Recall that *laten* has two main high density regions, which are close to the ones of *doen* and are related to several frequent lexically specific constructions (see Chapter 5). The high frequency of these patterns boosts the relative weight of their semantic and other features in the category structure. In contrast, the causation type with the affected Causer, which is the most distinctive sense of *laten*, is not represented by highly frequent low-level schemata, and its exemplars are more heterogeneous. This lack of lexical fixation, however, may suggest greater productivity of the sense in comparison with the other senses.

A similar lack of correlation between intercategorial predictions of membership (based on a classification algorithm) and intracategorial typicality ratings has been observed for the highly heterogeneous category of mammals by Ceulemans and Storms (2010) in their study of concrete lexical categories (mammals, birds, insects, fish, and reptiles/amphibians). Ceulemans and Storms hypothesize that typicality ratings of heterogeneous categories, such as mammals, can be biased towards one or more subcategories. They also demonstrate that the correlation improves significantly when the frequency of the specific category members is taken into account. To conclude, it seems that for semantically heterogeneous (polysemous) words and constructions typicality and cue validity may not be positively correlated, at least partly due to high familiarity/frequency of some of the senses and/or low-level schemata.

Although this and related issues need further investigation, including experimental evidence, the present analysis clearly demonstrates that we need to be very specific about the perspective and operationalization of salience effects. The same holds, in fact, for all semantic notions (cf. Stefanowitsch 2010). Unfortunately, the terms 'prototype' and 'salience' have been far too frequently misused in Cognitive Semantics.

The next step in developing corpus-based models of salience phenomena will involve refinement and experimental support of the straightforward measurements proposed here. The most important question is to what extent the highly frequent schemata, such as *doen denken aan*, contribute to the general constructional meaning of the parent category (*doen*). The cause of concern is their relative autonomy from the general schema. When they are reproduced and understood, the parent construction may not be fully activated (Bybee 2010: Ch. 3). At the same time, these patterns are more accessible to the speakers and therefore may serve as better category examples than the other members. It is important then to establish the effect of their entrenchment and autonomy on the intra- and intercategorial salience values of the exemplars.

Another consideration is whether other linguistic constructions related to causation and causality should be taken into account when creating the conceptual space of *doen* and *laten* and measuring the salience effects. There is no simple answer to this question. Causality is a fundamental aspect of cognition and language, and it permeates linguistic categorization at all levels (e.g. transitivity, modality, clausal connectives and prepositional marking). Altenberg (1984), for instance, made a list with 98 causal expressions in English, which includes various lexical and grammatical patterns. Whether the salience effects for the exemplars of *doen* and *laten* would change significantly if one added all possible causal expressions in Dutch, is a question for future research.

7.2. Is it done with *doen*?

In Section 2.3 of Chapter 2, two possible scenarios of the causative *doen*'s fate were discussed. According to the first one, *doen* has been shrinking as a category, and has lost not only the indirect causation meaning, but also a part of the direct causation semantics, which have been 'annexed' by *laten.*

The second one argues that the category contents in general (direct causation) has not changed, but the social experience that can be categorized as direct causation has reduced, hence the decrease in the use of *doen*, but no change in the use of *laten*. How can one interpret the results of the previous chapters in the context of this debate?

In general, *doen* occurs much less frequently than *laten* in all registers and varieties that I investigated. The results of my analyses support the results of Speelman and Geeraerts' quantitative study (2009), who found that *doen* is more frequently used in the Belgian variety and in more formal communication. These tendencies hold both for the ratios of *doen* vs. *laten* and the independent normalized frequencies of *doen* in the lectal samples. If we can use different varieties and registers of Dutch as a time machine, accepting the view that Belgian Dutch and formal registers have retained more archaic features than their counterparts (e.g. Speelman and Geeraerts 2009), then the results of the present study support the view of *doen* as a construction that has been losing its positions. The fact that *doen*, according to the Dutch CHILDES corpus, is not successfully learnt at an early age, whereas *laten* is well represented in terms of type and token frequencies already at the early stage, supports this conclusion. However, there is no outspoken geographic difference between the normalized frequencies of *laten*, except for the newspaper register. The construction is used more frequently in the Netherlandic newspapers than in the Belgian ones. Therefore, *doen* has been losing its positions, but there is no clear evidence that *laten* is taking over the former territory of *doen*.

The lectal differences in the semantic structure of the categories, as well as in the division of labour between the two constructions mainly reflect the variation in the referential situations, communicative functions, and frequencies of specific low-level schemata. The main conceptual distinction (directness or indirectness of causation) thus remains stable. It seems therefore that the main cause of the quantitative lectal differences is the difference in 'experience', not in 'concepts', according to the sociocognitive model of language variation proposed in Chapter 1 (Section 1.2).

However, the picture is more complex. One can find geographic differences in the organization of the semantic spaces of both constructions. First, the spaces are more fragmented in the Netherlandic data than in the Belgian registers. This fragmentation is explained by a higher degree of autonomy of the dense clusters with the collocations *doen denken aan* and *laten weten*, *laten zien*, etc., which are on average more prominent in the Netherlandic variety. These meanings are not prototypically causative from the conceptual point of view because they do not involve a change in the Causee or another entity as a result of the Causer's impingement. In addition, the causing and the caused events conveyed by these constructions are not easily separable because, as a rule, they are not observable directly. Note that in many other European languages, these senses are expressed as one lexeme, which indicates a tight integration of the events. Lexical alternatives also exist in Dutch, for instance, *tonen* “show”, which is used more frequently in Belgian Dutch than *laten zien* (according to the verb's frequencies in the newspapers and spoken data, this preference is statistically significant with $p < 0.001$).

Not surprisingly, the semantics of these autonomous schemata often goes beyond causation *per se*. As mentioned above, when using *doen denken aan* with an implicit Causee, the speaker sometimes refocuses on the properties of the Causer, and backgrounds the mental caused event. A similar refocusing also happens in the causeeless constructions with *laten zien* and *laten weten*, when the Causer's act of showing or informing becomes more important that the Causee's act of learning or perception. In fact, the expression *laten weten*, which occurs very frequently in the Netherlandic newspapers, is used as an evidentiality marker, which names the source of information reported by the journalist (cf. *aldus* “according to", which performs a similar function).

Thus, we have evidence of several processes, which might be going on, if we accept the view of the Netherlandic variety as the more 'advanced' one: lexicalization, subjectification, decausativization as a kind of semantic bleaching, and weakening of the parent schema due to the growing autonomy of some subschemata. Both *doen* and *laten* seem to be affected by these processes. Of course, one needs diachronic evidence to test this hypothesis.

For *doen*, which is already very restricted quantitatively and qualitatively in both national varieties, these processes may have more serious consequences than for *laten*. The very high relative frequency of *doen denken aan* can lead to a dramatic change of the entire category. Ultimately, *doen* can become a bound morpheme in the structures with several frequent Effected Predicates. On the other hand, if *laten* in general is also becoming less 'causative', we can also expect further semantic specialization and formal individualization of its subschemata, and/or a rise of alternative causative structures.

7.3. Towards an operationalizable semantic theory

Linguists are only beginning to develop tools to model different aspects of semantic and lectal variation. The present study proposes a number of techniques that can be used for this goal. But it is also important to interpret the results from a more general perspective. Has this modest analysis of only two constructions taught us anything about semantic and lectal variation in general?

First, the MDS analyses of the categories revealed cloud-like structures with a continuum between the most densely populated clusters. This fact supports the Semantic Map Connectivity Hypothesis by Croft (2001), which poses that constructional uses should be semantically connected – at least, historically. In other words, we should not normally expect the exemplars of one linguistic category to form an archipelago of semantic islands. This principle can be easily explained if we apply the notion of analogy: for novel occurrences to take place, they must be sufficiently similar to already existing ones (Bybee 2010: 57). The new exemplars will appear in the close vicinity of the old exemplars, which results in the continuum between the main senses.

It is also an interesting and somewhat unexpected finding that the main dimensions of the semantic space of both constructions are very similar: mental vs. non-mental events, and intentional vs. unintentional actions of the Causee. Moreover, these dimensions coincide with the fundamental distinctions between human actions (Malle 2005). In addition, the same dimensions were discovered for English causative constructions by Levshina et al. (In press) with the help of a different method. It would be interesting to test if the same or comparable dimensions would emerge if we test other predicative constructions in various languages. This would empirically support the existence of a universal conceptual space (cf. Croft 2001: 105).

The study of *doen* and *laten* also shows that the size of the category (in terms of its diversity) may correlate with the number of important semantic dimensions along which its semantic structure is organized and the meaning extensions are created. This hypothesis can be easily tested: the category size can be operationalized as the average distance between the exemplars, and the dimensions can be identified with the help of the variable-specific MDS maps, although it would be convenient to have more formal tools, which have yet to be developed.

Next, the results support the previous usage-based studies of frequency effects in language change and variation. For instance, it has been shown previously that the high token frequency of a constructional subschema increases the semantic autonomy of the latter from the parent schema (e.g. Bybee 2010: Ch.3). This is what we can see clearly on the semantic maps. The more populated the cluster, the more separated it is from the rest of the exemplars – i.e. it shares less features with them. This phenomenon is due to direct access to the entrenched subschema without (full) activation of the parent schema (Bybee 2010: 50). The method proposed in the present study offers a convenient intuitive way of detecting such effects visually. In addition, regression modelling with mixed effects can help in identification of the semantic idiosyncrasies that serve as evidence of relative semantic autonomy (see Section 6.3.2 of Chapter 6).

Another important effect of frequency, in Bybee's theory, is stability, which can be interpreted at the micro-level, in terms of individual categorization choices, and at the macro-level, as a factor in language change. The highly frequent Effected Predicates (*weten* "know", *zien* "see", *vallen* "fall", *denken aan* "think of" and some others) display a nearly exclusive preference for either *doen* or *laten* in the data. This can be interpreted as evidence of exemplar effects in constructional categorization (cf. Chapter 1, Section 1.1). From the historical perspective, if we accept the view of the causative *doen* as a gradually disappearing category, then we should expect it to be fossilized in the combinations with *denken aan* and several other predicates, such as *geloven* "believe" and *vermoeden* "suppose".

I hope that future works on the semantics of linguistic categories will benefit from the findings (or learn from the mistakes) of the present study in its theoretical and methodological aspects, and take new steps in this challenging but exciting enterprise. Much fine-tuning should be done, and converging evidence is needed, especially experimental support. The future will also tell whether my version of the story of *doen* is correct. It is with these hopes and expectations that I finish this chapter and this thesis.

Appendix 1. Local contextual variables

#	Variable	Values	Notes
Features of the Effected Predicate and the caused event			
1	*EPTrans* non-prepositional valency (transitivity)	- *EPTrans.Tr:* direct object (*Ik liet mijn huis schilderen.*) - *EPTrans.IntrIO:* indirect object (*U deed mij uw antwoord in het weekend toekomen.*) - *EPTrans.Ditr:* direct and indirect objects (*Ik laat me niks voorschrijven.*) - *EPTrans.IntrCop:* copula use (*Het deed hun briefwisseling authentiek lijken.*) -*EPTrans.Intr:* none of the above (*De politie liet hem gaan.*)	
2	*EPPrep* prepositional complements	- *EPPrep.aan*, *EPPrep.van*, etc. - *EPPrep.None.*	Criterion: unlike adverbial modifiers, prepositional complements cannot be omitted
3	*EP* Effected Predicate (lemma)	- *EP.denken* - *EP.zien* - *EP.voor_komen* - *EP.voorkomen,* etc.	In case of more than one EP, only the first one is considered.
4	*CausedSemS* source semantic domain of the caused event	- *CausedSemS.Phys:* physical (*De politie liet de auto stoppen.*) - *CausedSemS.Ment:* mental (*Het doet mij aan mijn ouders denken.*) - *CausedSemS.Soc:* social (*Hij liet zich verontschuldigen.*)	The event is classified according to its literal sense.
5	*CausedSemT* target semantic domain of the caused event	- *CausedSemT.Phys:* physical (*De politie liet de auto stoppen.*) - *CausedSemT.Ment:* mental (*Het doet een belletje rinkelen.*) - *CausedSemT.Soc:* social (*Hij liet zijn licht schijnen over het probleem.*)	The event is classified according to its figurative sense in case of figurative

			use, otherwise equal to *CausedSemS*
Features of the Causer			
6	*CrSynt* syntactic expression of the Causer	- *CrSynt.NP:* (pro)nominal phrase (*De politie liet hem gaan.*) - *CrSynt.NP:* clause (*Welke gevolgen dat heeft, laat zich raden.*) - *CrSynt.Impl:* implicit (*Laat eens weten waar het om gaat.*)	
7	*CrPOS* Causer's part of speech	- *CrPOS.Noun:* noun (*De politie liet hem gaan.*) - *CrPOS.Pron:* pronoun (*Hij liet zich niet verrassen.*) - *CrPOS.Verb:* substantivized verb (*Het opstellen van de verdediger deed veel stof opwaaien.*) - *CrPOS.Adj:* substantivized adjective (*Dat laatste doet veel vragen rijzen*) - *CrPOS.Num:* numeral (*2 doet me denken aan Moby.*)	Substantivization was coded if the substantivized unit was not registered in dictionaries as a noun.
8	*CrPers* Causer's grammatic person	- *CrPers.1st* (*Ik liet me niet verrassen.*) - *CrPers.2nd* (*Je liet je niet verrassen.)* - *CrPers.3rd* (*Hij liet zich niet verrassen.*)	
9	*CrNo* Causer's grammatic number	- *CrNo.Sg:* singular (*Hij liet zich niet verrassen.*) - *CrNo.Pl:* plural (*De proteststemmen lieten van zich horen.*)	
10	*CrDef* Causer's definiteness	- *CrDef.Def:* definite (*De dirigent laat het orkest kleurrijk klinken.*) - *CrDef.Indef:* indefinite (*Een winstwaarschuwing deed de koers dalen.*)	"NA" if not applicable or the article is omitted in a header
11	*CrSem* semantic class of the Causer	- *CrSem.Hum:* human individual(s) (*De minister liet weten dat...*) - *CrSem.Org:* organization (*De NAVO liet weten dat...*) - *CrSem.HumUndef:* human undefined (individual or organization) (*Het is fout om hen te laten boeten.*) - *CrSem.Zoo:* animal (*De hond doet*	If the Causer is implicit, the class is determined with the help of the context.

		denken aan...) - *CrSem.Body:* body part (*Zijn ogen doen me denken aan...*) - *CrSem.Mech:* mechanism (*Deze auto doet denken aan...*) - *CrSem.CarryInfo:* material carrier of information (*De cd doet denken aan...*) - *CrSem.MatObj:* other material object or substance (*De tafel doet denken aan...*) - *CrSem.Abstr:* abstract entity (*Je antwoord doet denken aan..*)	
Features of the Causee			
12	*CeSynt* syntactic expression of the Causee	- *CeSynt.NP:* zero-marked (pro)nominal phrase (*De gelijkmaker deed de wedstrijd kantelen.*) - *CeSynt.Door:* NP marked with *door* (*Hij liet zich door de reacties afschrikken.*) - *CeSynt.Aan:* NP marked with *aan* (*Het concern liet dat weten aan de aandeelhouders.*) - *CeSynt.Van.* NP marked with *van* (if substantivization) (*Het laten vallen van de club zou raadzetels kosten.*) - *CeSynt.Clause:* clause (*Wat ik ga doen, laat ik van het moment afhangen.*) - *CeSynt.Impl:* implicit (*Hij liet zich verrassen.*)	
13	*CePOS* Causee's part of speech	- *CePOS.Noun:* noun (*De gelijkmaker deed de wedstrijd kantelen.*) - *CePOS.Pron:* pronoun (*Ik liet me gaan.*) - *CePOS.Verb:* substantivized verb (*Ik liet het downloaden beginnen.*) - *CePOS.Adj:* substantivized adjective (*Hij liet paars-wit floreren.*) - *CePOS.Num:* numeral (*Ik moet er vijf of zes laten afvallen.*)	Substantivization was coded if the substantivized unit was not registered in dictionaries as a noun.
14	*CePers* Causee's grammatic person	- *CePers.1st* (*Ik liet me gaan.*) - *CePers.2nd* (*Je liet je gaan.*) - *CePers.3rd* (*Hij liet zich gaan.*)	
15	*CeNo* Causee's	- *CeNo.Sg:* singular (*Ik liet me gaan.*) - *CeNo.Pl:* plural (*We lieten ons gaan.*)	

	grammatic number		
16	*CeDef* Causee's definiteness	- *CeDef.Def:* definite (*Hij liet zijn sigaret vallen.*) - *CeDef.Indef:* indefinite (*Hij liet een sigaret vallen.*)	
17	*CeSem* semantic class of the Causee	- *CeSem.Hum:* human individual(s) (*De politie liet de dader ontsnappen.*) - *CeSem.Org:* organization (*Dat deed de bank de verwachte winst verhogen.*) - *CeSem.HumUndef:* undefined human (individual or organization) (*We lieten ons huis schilderen.*) - *CeSem.Zoo:* animal (*De eigenaar liet de hond creperen.*) - *CeSem.Body:* body part (*Die muziek doet harten sneller slaan.*) - *CeSem.Mech:* mechanism (*Hij liet de motor draaien.*) - *CeSem.CarryInfo:* material carrier of information (*De zangeres deed twee cd's het licht zien.*) - *CeSem.MatObj:* other material object or substance (*Het was de druppel die de emmer deed overlopen.*) - *CeSem.Abstr:* abstract entity (*Hij liet het plan varen.*)	If the Causee is implicit, the class is determined with the help of the context.
18	*CeIntent* whether the Causee performs the caused event intentionally	- *CeIntent.Yes*: intentional action (*Ik liet hen mijn huis schilderen.*) - *CeIntent.No*: unintentional action or process(*Hij liet de auto stoppen.*) - *CeIntent.Undef:* the event is difficult to interpret as clearly intentional or unintentional (*Het doet het ergste vermoeden.*)	Several tests, e.g. "..., *and the Causee did that because he/she wanted to."*
19	*CeRole* role of the Causee in the event specified in the Effected Predicate	- *CeRole.Cause*: the Causee causes a change in another entity (*Ik liet mijn huis schilderen.*); - *CeRole.Change*: the Causee undergoes a change (*Hij liet de auto stoppen.*); - *CeRole.Oth*: no transfer of energy or change (*Ze liet me haar nieuw book lezen.*)	

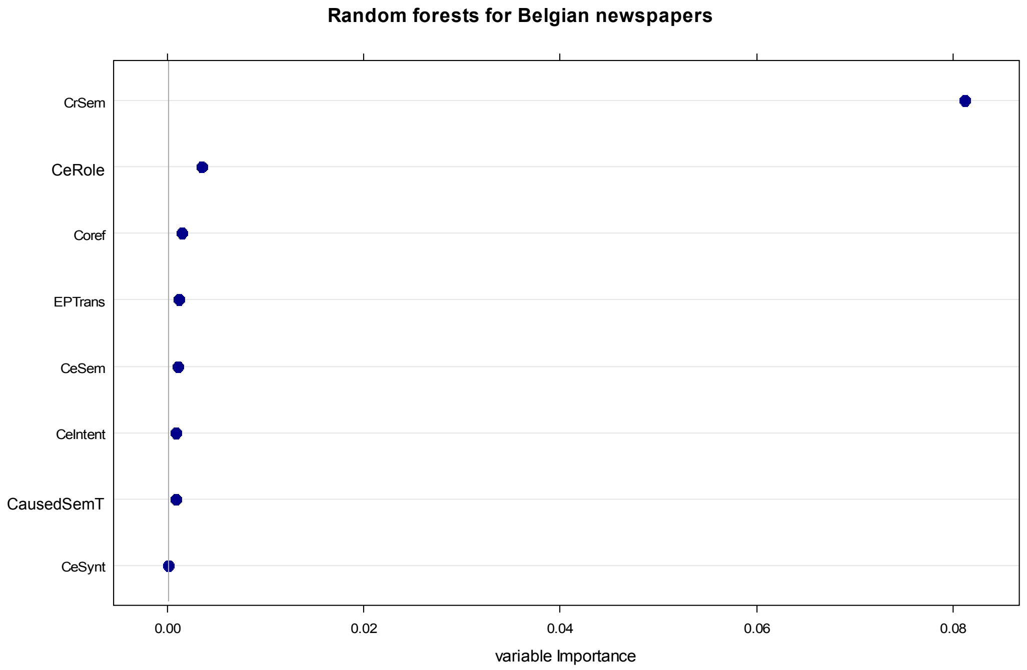
Features of the Affectee			
20	*AffSynt* syntactic expression of the Affectee	- *AffSynt.NP:* (pro)nominal phrase (*Ik liet mijn huis schilderen.*) - *AffSynt.Clause:* clause (*Dat doet vermoeden dat de problemen dieper zitten.*)	
21	*AffPOS* Affectee's part of speech	- *AffPOS.Noun:* noun (*Ik liet mijn huis schilderen.*) - *AffPOS.Pron:* pronoun (*Hij liet zich verrassen.*) - *AffPOS.Verb:* substantivized verb (*Dat doet denken aan het spuiten van geklopt eiwit.*) - *AffPOS.Adj:* substantivized adjective (*Dat doet het ergste vermoeden.*)	Substantivization was coded if the substantivized unit was not registered in dictionaries as a noun.
22	*AffPers* Affectee's grammatic person	- *AffPers.1st* (*Ik liet me verrassen.*) - *AffPers.2nd* (*Je liet je verrassen.*) - *AffPers.3rd* (*Hij liet zich verrassen.*)	
23	*AffNo* Affectee's grammatic number	- *AffNo.Sg:* singular (*Ik liet me verrassen.*) - *AffNo.Pl:* plural (*We lieten ons verrassen.*)	
24	*AffDef* Affectee's definiteness	- *AffDef.Def:* definite (*Ik liet mijn huis schilderen.*) - *AffDef.Indef:* indefinite (*Laat iets weten!*)	
25	*AffSem* semantic class of the Affectee	- *AffSem.Hum:* human individual(s) (*Ik liet me verrassen.*) - *AffSem.Org:* organization (*Ajax liet zich meeslepen in een vlaag van euforie.*) - *AffSem.HumUndef:* human undefined (individual or organization) (*Men moet zich laten registreren.*) - *AffSem.Zoo:* animal (*Ze lieten hun hond inslapen.*) - *AffSem.Body:* body part (*Ze laten hun haar blonderen.*) - *AffSem.Mech:* mechanism (*Hij liet 60*	

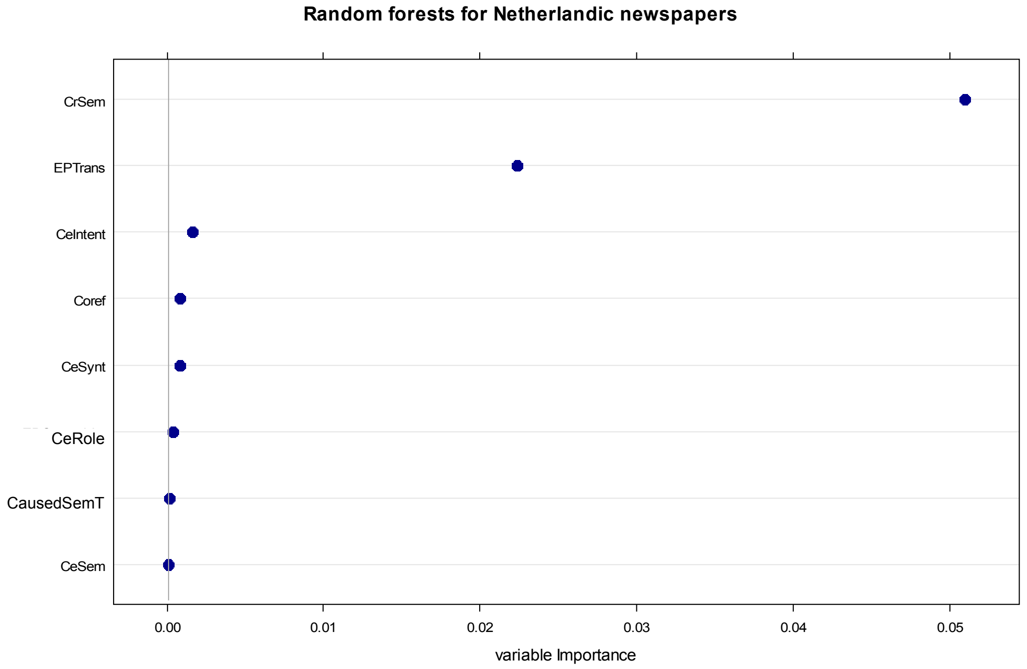
		limo's overvliegen.) - *AffSem.CarryInfo:* material carrier of information (*Hij liet een serie platen maken.*) - *AffSem.MatObj:* other material object or substance *Ik liet mijn huis schilderen.*) - *AffSem.Abstr:* abstract entity (*Dat doet het ergste vermoeden.*)	
Relationships between the main participants			
26	*Coref* coreferentiality of the Causer with the other participants	- *Coref.CrCe:* Causer = Causee (*Ik liet me gaan.*) - *Coref.CrAff:* Causer= Affectee (*Hij liet zich verrassen.*) - *Coref.CeAff:* Causee = Affectee (*Ik liet hem zich wassen.*)	Reflexive pronouns are used as indicators.
27	*Possess* possession relationships between the Causer and the other participants, in a broad sense	- *Possess.CrCe:* Causer 'owns' Causee (*Hij liet zijn sigaret vallen.*) - *Possess.CrAff:* Causer 'owns' Affectee (*Ik liet mijn huis schilderen.*) - *Possess.CeAff:* Causee 'owns' Affectee (*Ik liet hem zijn liedjes zingen.*)	Possessive pronouns or case are used as formal indicators.
Other features of the Causative Construction			
28	*SyntFun* syntactic function of the construstion	- *SyntFun.Pred:* predicate (*De politie liet hem gaan.*) - *SyntFun.Inf:* infinitival clause (*Hij deed dat om te laten zien dat...*) - *SyntFun.Subst:* subject or object – only for substantivized auxiliaries (*Het dubbele laten produceren, is corrupt.*)	
29	*Modal* modal verbs modifying the auxiliary	- *Modal.kunnen* (Ik *kan je dat laten zien.*) - *Modal.moeten* - *Modal.mogen* - *Modal.willen* - *Modal.hoeven*	

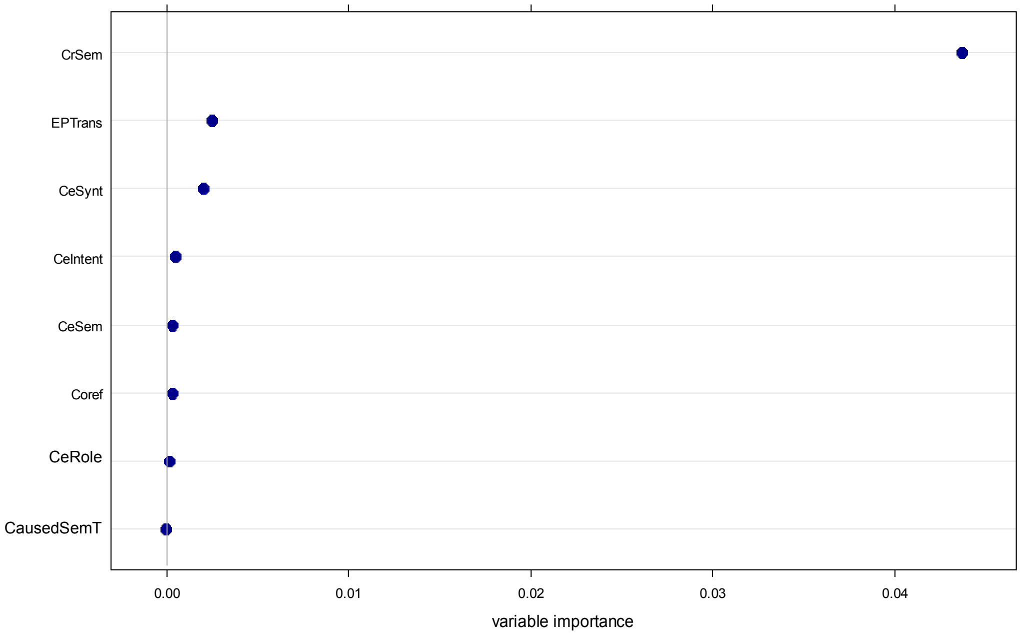
		- *Modal.durven* - *Modal.None*	
30	*Adv* adverbial modifiers	- *Adv.Manner:* manner and means (*langzaam, zodoende*) - *Adv.Place* (*in Amsterdam*) - *Adv.Time* (*gisteren*) - *Adv.Degree* (*een beetje*) - *Adv.Dur:* duration (*twee dagen*) - *Adv.Freq:* frequency (*altijd*) - *Adv.Rep:* repetition (*weer*) - *Adv.Oth:* other (none of the above, or several different types together in one context) - *Adv.None:* none	
31	*Neg* negation	- *Neg.Cr:* negation of the Causer (*Geen enkele persconferentie kan de rust doen weerkeren.*) - *Neg.Ce:* negation of the Causee (*Zijn naam doet geen belletje rinkelen.*) - *Neg.Aff:* negation of the Affectee (*De komst van de nieuwe trainer deed geen beterschap vermoeden.*) - *Neg:* other types of negation in the construction (usually negation with *niet*) (*Dat deed de verkoop niet stijgen.*) - *Neg.Clause:* construction is subordinated to or a part of a negative clause (*Dat is geen argument om de tv-zenders meer te doen betalen.*) - *Neg.Pos*: no negation	
General features of the clause			
32	*ClauseMood* grammatical mood of the clause	- *ClauseMood.Ind:* indicative (*De politie liet de auto stoppen.*) - *ClauseMood.Imp:* imperative (*Laat iets weten!*) - *ClauseMood.Subj:* "subjunctive" (irrealis) (*Als ik veel geld had, zou ik een groot huis laten bouwen.*)	The "Subjunctive" is determined semantically.
33	*ClauseTense* grammatical tense of the clause	- *ClauseTense.Pres:* presens (*Ik laat hem gaan.*) - *ClauseTense.Past:* imperfectum (*Ik liet hem gaan.*) - *ClauseTense.PrPerf:* perfectum (*Ik heb hem laten gaan.*)	

		- *ClauseTense.PastPerf:* plusquamperfectum (*Ik had hem laten gaan.*) - *ClauseTense.Fut:* futurum (*Ik zal hem laten gaan.*) - *ClauseTense.FutPast:* futurum praereriti (*Ik zou hem laten gaan.*)	
34	*Clause* type of clause where the causative is found	- *Clause.Main:* main (*Het deed me het ergste vermoeden.*) - *Clause.Rel:* relative (*Haar reactie, die me het ergste deed vermoeden,...*) - *Clause.Comp:* complement (*Ik zei dat haar reactie mij het ergste deed vermoeden*) - *Clause.Adv:* adverbial (*Ik zei dat omdat haar woorden mij het ergste deden vermoeden.*) - *Clause.Add:* 'additional' (*Ze ging weg, wat me het ergste deed vermoeden.*)	Clauses introduced with *want* are treated as subordinate adverbial clauses.
General features of the sentence (utterance)			
35	*Sent* sentence type according to its communicative function	- *Sent.Decl:* declarative (*Ik liet hem gaan.*) - *Sent.Imp:* imperative (*Let wel, degenen die een fancard laten aanmaken!*) - *Sent.Q:* interrogative (*Moet ik hem laten gaan?*)	

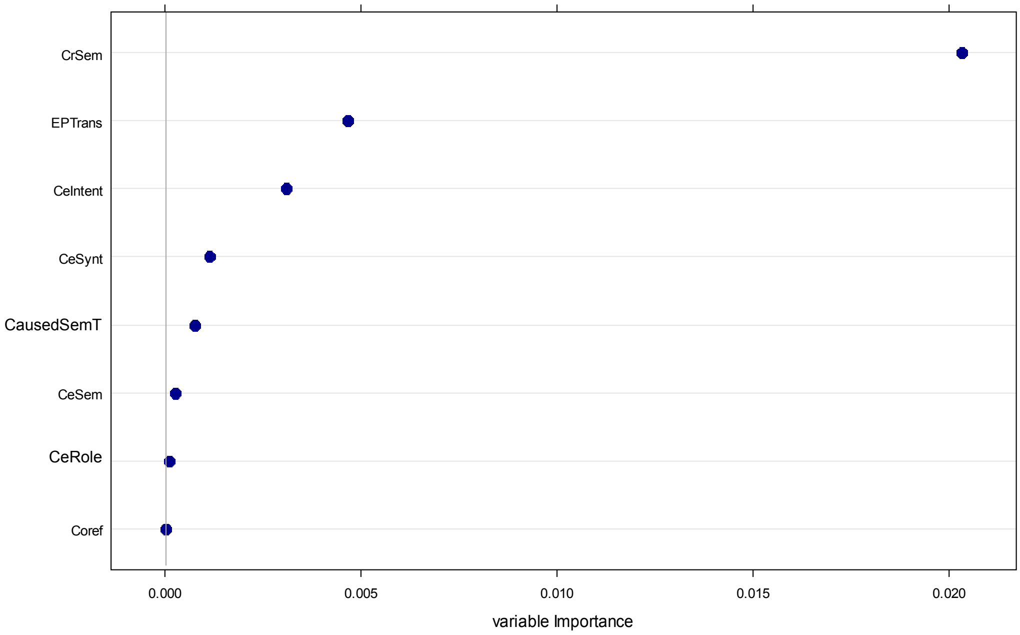
Appendix 2. Random forests for six lects



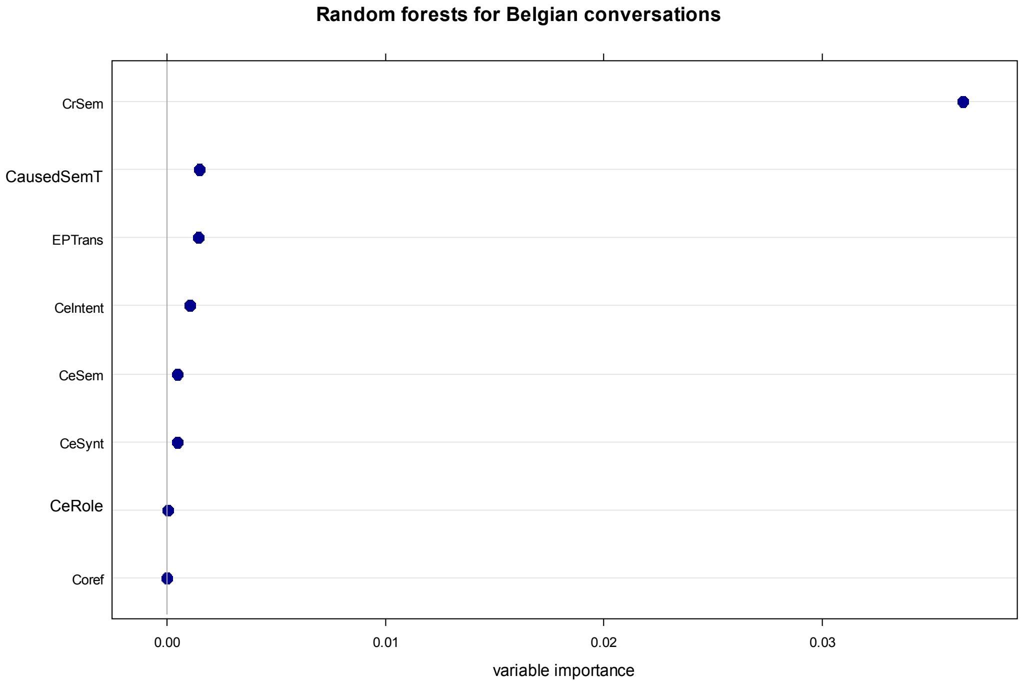


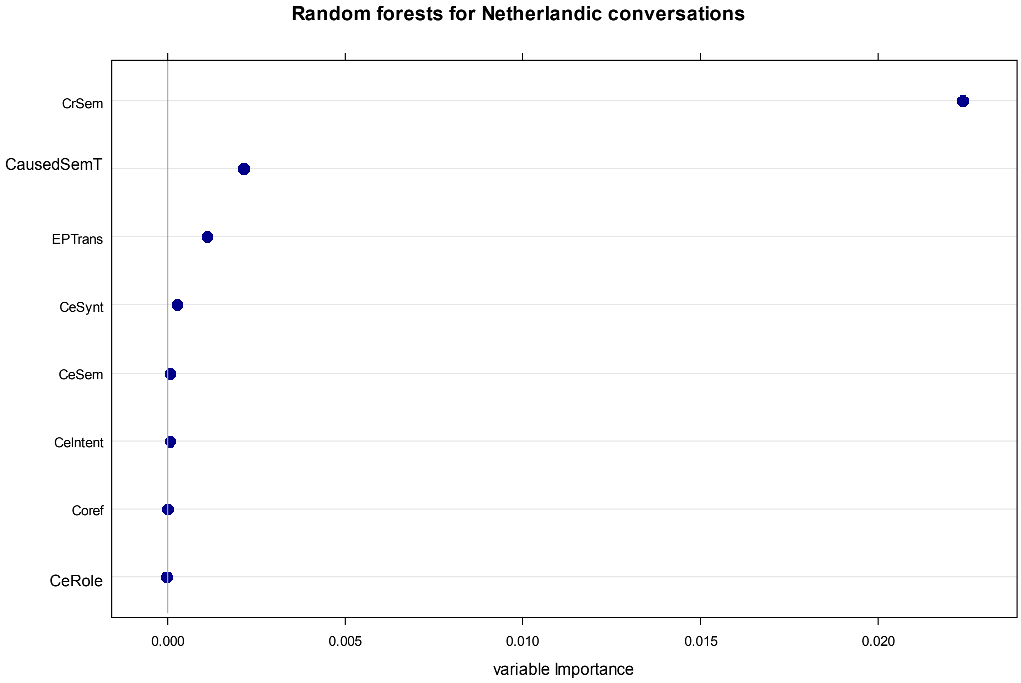


Random forests for Belgian Usenet



Random forests for Netherlandic Usenet


Random forests for Belgian conversations
CrSem
CausedSemT
EPTrans
CeIntent
CeSem
CeSynt
CeRole
Coref
0.00
0.01
0.02
0.03
variable importance


Random forests for Netherlandic conversations
CrSem
CausedSemT
EPTrans
CeSynt
CeSem
CeIntent
Coref
CeRole
0.000
0.005
0.010
0.015
0.020
variable Importance

References

Abbot-Smith, Kirsten & Michael Tomasello. 2006. Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review* 23. 275–290.

Achard, Michel. 2002. Causation, constructions, and language ecology: An example from French. In Masayoshi Shibatani (ed.), *The grammar of causation and interpersonal manipulation*, 127–155. Amsterdam/Philadelphia: John Benjamins.

Aguinis, Herman, 2004. Regression analysis for categorical moderators. Guilford: New York.

Altenberg, Bengt. 1984. Causal linking in spoken and written English. *Studia Linguistica* 38. 20–69.

ANS 1997. Haeseryn, Walter, Kirsten Romijn, Guido Geerts, Jaap de Rooij & Maarte C. van den Toorn. *Algemene Nederlandse Spraakkunst* [Standard Dutch Grammar]. 2nd edn. Groningen/Deurne: Martinus Nijhoff/Wolters Plantyn.

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert and Arne Zeschel. 2010. Cognitive Corpus Linguistics: five points of debate on current theory and methodology. *Corpora* 5 (1). 1–27.

Arppe, Antti & Juhani Järvikivi. 2007. Every method counts: combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3 (2). 131–159.

Atkins, Beryl T. S. 1987. Semantic ID tags: Corpus evidence for dictionary senses. In *The uses of large text databases*. *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary*, University of Waterloo, 17–36. Waterloo, Canada.

Auer, Peter. 2005. Europe' Sociolinguistic unity, or: A typology of European dialect/standard constellations. In Nicole Delbecque, Johan van der Auwera and Dirk Geeraerts (eds.), *Perspectives on variation: Sociolinguistic, historical, comparative*, 8-42. Berlin/New York: Mouton de Gruyter.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Barlow, Michael & Suzanne Kemmer (eds.). 2000. *Usage based models of language*. Stanford: CSLI Publications.

Baroni, Marco & Alessandro Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics* 36 (4). 673–721.

Bergen, Benjamin. 2007. Experimental methods for simulation semantics. In Monica Gonzalez-Marquez, Irene Mittleberg, Seanna Coulson & Michael J. Spivey (eds.), *Methods in Cognitive Linguistics*, 277-301. Amsterdam/Philadelphia: John Benjamins.

Berlin, Brent. 1978. Ethnobiological classification. In Eleanor Rosch & Barbara B. Lloyd (eds), *Cognition and categorization*, 9–26. Hillsdale, NJ: Erlbaum.

Berthele, Raphael. 2010. Investigation into the folk's mental models of linguistic varieties. In *Recent Advances in Cognitive Sociolinguistics*, 265–290. Berlin/New York: Mouton de Gruyter.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Blank, Anderas. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In Andreas Blank & Peter Koch (eds.), *Historical semantics and cognition*, 61–89. Berlin/New York: Mouton de Gruyter.

Boon, Ton den & Geeraerts, Dirk (eds.). 2005. *Van Dale: Groot Woordenboek der Nederlandse Taal* [Van Dale: Large Dictionary of the Dutch language]. 14th edn. Utrecht: Van Dale Lexicografie.

Borg, Ingwer & Patrick Groenen. 1997. *Modern multidimensional scaling: theory and applications*. New York: Springer.

Bouma, Gosse, Gertjan van Noord & Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. Walter Daelemans, Khalil Simia'an, Jorn Veenstra & Jakub Zavrel (eds.), *Computational linguistics in the Netherlands 2000: Selected papers from the 11th CLIN meeting*, 45–59. Amsterdam: Rodopi.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Boume, Irene Krämer, & Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan & Marilyn Ford, 2010. Predicting Syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1): 168–213.

Bresnan, Joan & Jennifer Hay. 2008. Gradient Grammar: An Effect of Animacy on the Syntax of *give* in New Zealand and American English. *Lingua* 118(2): 245–59.

Brugman, Claudia M. 1983. *The story of over.* Bloomington: Indiana University. Linguistics club.

Bybee, Joan L. 1985. *Morphology: a study of the relation between meaning and form.* Amsterdam/Philadelphia: John Benjamins.

Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 529–551.

Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, Joan L.& David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82(2). 323–355.

Ceulemans, Eva & Gert Storms. 2010. Detecting intra and inter categorical structure in semantic concepts using HICLAS. *Acta Psychologica* 133, 296–304.

Clark, Herbert H. 1999. On the origins of conversation. *Verbum* 21,147–161.

Colleman, Timothy. 2010. Beyond the dative alternation: The semantics of the Dutch *aan*-Dative. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, 271–303. Berlin/New York: De Gruyter Mouton.

Coopmans, Peter & Martin Everaert. 1988. The simplex structure of complex idioms: The morphological status of *laten*. In Martin Everaert, Arnold Evers, Riny Huybregts & Mieke Trommelen (eds.), *Morphology and modularity. In honour of Henk Schultink*, 95–104. Dordrecht: Foris.

Croft, William. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Croft, William. 2009. Toward a social cognitive linguistics. In In Vyvyan Eans and Stéphanie Pourcel (eds.), *New Directions in Cognitive Linguistics*, 395–420. Amsterdam/Philadelphia: John Benjamins.

Croft, William & D. Allan Cruse. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Croft, William & Keith T. Poole. 2008. Inferring universals from grammatical variation: multidimensional scaling for typological analysis. *Theoretical Linguistics* 34.1–37.

Dąbrowska, Ewa. 2009. Words as constructions. In Vyvyan Eans and Stéphanie Pourcel (eds.), *New Directions in Cognitive Linguistics*, 201–223. Amsterdam/Philadelphia: John Benjamins.

D'Andrade, Roy G. 1987. A folk model of the mind. In Dorothy Holland and Naomi Quinn (eds.), *Cultural models in language and thought*, 112–148. Cambridge: Cambridge Univeristy Press.

Davidse, Kristin & Liesbet Heyvaert. 2003. On the so-called 'middle' construction in English and Dutch. In Sylviane Granger, Jacques Lerot & Stephanie Petch-Tyson (eds.), *Corpus-based approaches to contrastive linguistics and translation studies*, 57–73. Amsterdam/New York: Rodopi.

De Sutter, Gert. 2009. Towards a multivariate model of grammar: The case of word order variation in Dutch clause final verb clusters. In Andreas Dufter, Jürg Fleischer & Guido Seiler (eds.), Describing and modeling variation in grammar, 225–254. Berlin/New York: Mouton de Gruyter.

Degand, Liesbeth. 1996. Causation in Dutch and French: Interpersonal aspects. In Ruqaiya Hasan, Carmel Cloran & David Butt (eds.), *Functional Descriptions: Theory in Practice*, 207–235. Amsterdam/Philadelphia: John Benjamins.

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

Dik, Simon C. 1980. *Studies in functional grammar.* London: Academic Press.

Divjak, Dagmar. 2006. Ways of intending: A corpus-based cognitive linguistic approach to near synonyms in Russian. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in Cognitive Linguistics. Corpus-based approaches to syntax and lexis*, 19–56. Berlin/New York: Mouton de Gruyter.

Divjak, Dagmar. 2010a. Corpus-based evidence for an idiosyncratic aspect-modality relation in Russian. In In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, 305–330. Berlin/New York: De Gruyter Mouton.

Divjak, Dagmar. 2010b. *Structuring the lexicon: A clustered model for* *near-synonymy*. Berlin/New York: Mouton de Gruyter.

Divjak, Dagmar & Stefan Th. Gries. 2008. Clusters in the mind? Converging evidence from near-synonymy in Russian. *The Mental Lexicon* 3: 59–81.

Draye, Luk. 1998. The case of the causee: on the competition between dative and accusative in German *lassen* and in Dutch *laten*-constructions. In Willy Van Langendonck & William Van Belle (eds.), *The dative. 2: Theoretical and contrastive studies*, 75–111. Amsterdam/Philadelphia: John Benjamins.

Duinhoven, Anton M. 1994a. Doen en laten in beweging [Doen and laten in motion]. *Forum der Letteren* 35 (4). 272–276.

Duinhoven, Anton M. 1994b. Het hulpwerkwoord doen heeft afgedaan [The auxiliary verb doen has had its day]. *Forum der Letteren* 35 (2). 110–131.

Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4). 453–476.

Firth, John R. 1957. A synopsis of linguistic theory 1930–1950. In John R. Firth (ed.), *Studies in linguistic analysis*, 1–32. Oxford: Philological Society.

Estes, William K. *Classification and cognition*. New York/Oxford: Oxford Univeristy Press/Claredon Press.

Everitt, Brian S., Sabine Landau & Morven Leese. 2001. *Cluster analysis*. London: Arnold.

Geeraerts, Dirk. 1998. The semantic structure of the indirect object in Dutch. In Willy Van Langendonck & William Van Belle (eds.), *The Dative*. Vol. 2. *Theoretical and contrastive studies*, 1985–210. Amsterdam/Philadelphia: John Benjamins.

Geeraerts, Dirk. 2005. Lectal variation and empirical data in Cognitive Linguistics. In Ruiz de Mendoza Ibáñez, Francisco J. & Peña Cervel, M. Sandra (eds.), *Cognitive Linguistics: internal* *dynamics and interdisciplinary interaction*, 163–189. Berlin: Mouton de Gruyter.

Geeraerts, Dirk. 2006 [2000]. Salience phenomena in the lexicon. A typology. In Dirk Geeraerts, *Words and other wonders. Papers on lexical and semantic topics*, 74–97. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk. 2010a. Lexical variation in space. In Peter Auer, P. & Jürgen E. Schmidt (eds.), *Language and Space*. Vol. I: *Theories and Methods*, 820–836. Mouton de Gruyter, Berlin.

Geeraerts, Dirk. 2010b. The doctor and the semantician. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, 63–78. Berlin/New York: De Gruyter Mouton.

Geeraerts, Dirk. 2010c. *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Geeraerts, Dirk & Stefan Grondelaers. 1995. Looking back at anger: Cultural traditions and metaphorical patterns. In John R. Taylor & Robert E. MacLaury (eds.), *Language and the Cognitive Construal of the World*, 153-179. Berlin/New York: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers & Peter Bakema. 1994. *The structure of lexical variation. Meaning, naming, and context*. 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.*[Convergence and divergence in the Dutch lexicon. A study of clothing and football terms]. Amsterdam: Meertens Instituut.

Geeraerts, Dirk, Gitte Kristiansen & Yves Peirsman (eds.). 2010. *Advances in Cognitive Sociolinguistics*. Berlin/New York: Mouton de

Gruyter.

Gilquin, Gaëtanelle. 2006. The place of prototypicality in corpus linguistics: Causation in the hot seat. In Stephan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, 159–191. Berlin: Mouton de Gruyter.

Gilquin, Gaëtanelle. 2010. *Corpus, cognition and causative constructions*. Amsterdam/Philadelphia: John Benjamins.

Glynn, Dylan. 2007. *Mapping meaning. Towards a usage-based methodology in Cognitive Semantics*. Leuven: Catholic Univeristy of Leuven dissertation.

Goldberg, Adele E. 1995. *Constructions. A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2002. Surface generalizations: An alternative to alternations. *Cognitive Linguistics* 13. 327–356.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties, *Biometrics* 27. 857–874.

Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: a study of Particle Placement*. London/New York: Continuum Press.

Gries, Stefan Th. 2006. Corpus-based methods and Cognitive Semantics: The many senses of *to run*. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in Cognitive Linguistics. Corpus-based approaches to syntax and lexis*, 57–99. Berlin/New York: Mouton de Gruyter.

Gries, Stefan Th. & Dagmar Divjak. 2010. Quantitative approaches in usage-based Cognitive Semantics: Myths, erroneous assumptions, and a proposal. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, 333–353. Berlin/New York: De Gruyter Mouton.

Gries, Stefan Th. & Naoki Otani. 2010. Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal* 34. 121–150.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9 (1). 97–129.

Gries, Stefan Th. & Anatol Stefanowitsch. 2010. Cluster analysis and the identification of collexeme classes. In John Newman & Sally Rice (eds.), *Empirical and experimental methods in cognitive/functional research*, 73–90. Stanford: CSLI Publications.

Grondelaers, Stefan, Dirk Speelman & Dirk Geeraerts 2002. Regressing on *er.* Statistical analysis of texts and language variation. In Annie Morin & Pascale Sébillot (eds.), *Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data,* 335–346. Rennes: Institut National de Recherche en Informatique et en Automatique.

Hanks, Patrick. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1 (1). 75–98.

Heylen, Kris. 2005. A quantitative corpus study of German word order variation. In Stephan Kepser & Magda Reis (eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives*, 241–264. Berlin: Mouton de Gruyter.

Heylen, Kris, Yves Peirsman & Dirk Geeraerts. 2008. Modelling Word Similarity. An Evaluation of Automatic Synonymy Extraction Algorithms. In N. Calzolari et al. (eds.), *Proceedings of the Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco, 28–30 May 2008, 3243–3249. Marrakech:

European language resources association.

Hilpert, Martin. 2010. The force dynamics of English complement clauses: A Collostructional Analysis. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, 155–178. Berlin/New York: De Gruyter Mouton.

Hilpert, Martin. Submitted. Visualizing language change with diachronic corpus data: Introducing the flipbook technique. Submitted to *International Journal of Corpus Linguistics*.

Hofstede, Geert H. 2001. *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. 2nd edn. Thousand Oaks: Sage.

Hopper, Paul J. and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56(2): 251–299.

van der Horst, Joop M. 1998. *Doen* in Old and Early Middle Dutch: A comparative approach. In Ingrid Tieken-Boon van Ostade, Marijke van der Wal and Arjan van Leuvensteijn (eds.), *'Do' in English, Dutch and German. History and present-day variation*, 53–64. Münster: Nodus Publicationen.

van der Horst, Joop M. 2008. *Geschiedenis van de Nederlandse syntaxis* [History of the Dutch syntax]. Vol.1. Leuven: Universitaire pers Leuven.

Hosmer, David W. & Stanley Lemeshow. 2000. *Applied Logistic Regression.* New York: Wiley.

Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased Recursive Partitionaing: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15(3): 651–674.

Keller, Rudi. 1994. *On language change: The invisible hand in language*. London & New York: Routledge.

Kemmer, Suzanne & Arie Verhagen. 1994. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5. 115–156.

Kristiansen, Gitte & René Dirven (eds.). 2008. *Cognitive Sociolinguistics: Language variation, cultural models and social systems*. Berlin/New York: Mouton de Gruyter.

Labov, William. 1971. The study of language in its social context. In Joshua A. Fishman (ed.), *Advances in the sociology of language*, Vol. 1, 152–216. The Hague: Mouton.

Lakoff, George. 1987. Women, fire and dangerous things: What categories reveal about the mind. Chicago: Univeristy of Chicago Press.

Lakoff, George. 1990. The invariance hypothesis: Is abstract reason based on image-schemas? *Cognitive Linguistics* 1(1), 39–74.

Landauer, Thomas K. & Susan T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104, 211–240.

Landsbergen, Frank, Robert Lachlan, Carel Ten Cate & Arie Verhagen. 2010. A cultural evolutionary model of patterns in semantic change. *Linguistics* 48 (2), 365–390.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar*. Vol. 1. *Theoretical prerequisites*. Stanford: Stanford Univeristy Press.

Langacker, Ronald W. 2005. Construction grammars: Cognitive, radical and less so. In Ruiz de Mendoza Ibáñez, Francisco J. & Peña Cervel, M. Sandra (eds.), *Cognitive Linguistics: internal dynamics and interdisciplinary interaction*, 101–159. Berlin: Mouton de Gruyter.

de Leeuw, Jan & Patrick Mair. 2009. Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software* 31(3). 1–30. http://www.jstatsoft.org/v31/i03/. (Last access 19.08.2011)

Levin, Beth. 1993. *English verb classes and alternations: a preliminary investigation*. Chicago: University of Chicago Press.

Levinson, Stephen C., & Sérgio Meira. 2003. 'Natural concepts' in the spatial topological domain - adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language* 79(3), 485–516.

Levshina, Natalia, Dirk Geeraerts & Dirk Speelman. 2011. Changing the world vs. changing the mind: Distinctive collexeme analysis of the causative construction with *doen* in Belgian and Netherlandic Dutch, 111-123. In F. Gregersen, J. Parrot & P. Quist (eds.), *Language variation - European perspectives III*. Selected papers from the 5th International Conference on Language Variation in Europe, Copenhagen, June 2009. Amsterdam: John Benjamins.

Levshina, Natalia, Dirk Geeraerts & Dirk Speelman. In press. Mapping constructional spaces: A contrastive study of English and Dutch analytic causatives. To appear in *Linguistics*.

Levshina, Natalia, Dirk Geeraerts & Dirk Speelman. Submitted. Towards a 3D-grammar. Variation of Dutch causative constructions. Submitted to *Journal of Pragmatics*.

Levshina, Natalia & Kris Heylen. In preparation. Data-driven Construction Grammar: A case of Dutch causative constructions.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, Montreal, Canada, August 1998, 768–774.

Loewenthal, Judith. 2003. Meaning and use of causeeless causative constructions with laten in Dutch. In Arie Verhagen & Jeroen van de Weijer (eds.), *Usage-Based Approaches to Dutch*, 97–130. Utrecht: LOT.

Lund, Kevin and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instrumentation, and Computers* 28(2). 203–208.

Maechler, Martin, Peter Rousseeuw, Anja Struyf & Mia Hubert. 2005. Cluster Analysis Basics and Extensions; unpublished.

Majid, Asifa, Marianne Gullberg, Miriam van Staden and Melissa Bowerman. 2007. How similar are semantic categories in closely related langugages? A comparison of cutting and breaking in four Germanic languages. *Cognitive Linguistics* 18 (2). 179–194.

Malle, Bertram F. 2005. Folk theory of mind: Conceptual foundations of social cognition. In Hassin, Ran R., James S. Uleman, & John A. Bargh (Eds.), *The New Unconscious*, 225–255. New York: Oxford University Press.

Medin, Douglas L. & Schaffer, Marguerite M. 1978. Context theory of classification learning. *Psychological Review* 85. 207-238.

Murphy, Gregory L. 2002. *The big book of concepts*. Cambridge, MA: MIT Press.

Nosofsky, Robert M. 1986. Attention, similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General* 115(1). 39–57.

Newman, John. 2010. Balancing acts: Empirical pursuits in Cognitive Linguistics. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, 79–99. Berlin/New York: De Gruyter Mouton.

Oostdijk, Nelleke H.J. 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.

Ordelman, Roeland, Franciska de Jong, Arjan van Hessen & Henri Hondorp. 2007. TwNC: a multifaceted Dutch News Corpus. *ELRA Newsletter* 12 (3–4), http://doc.utwente.nl/68090/ (28 May 2011)

Peirsman, Yves. 2010. *Crossing Corpora. Modelling Semantic Similarity across Languages and Lects*. Leuven: Catholic Univeristy of Leuven dissertation.

Peirsman, Yves, Kris Heylen & Dirk Geeraerts. 2010. Applying Word Space Models to Sociolinguistics. Religion Names before and after 9/11. In Dirk Geeraerts, Gitte Kristiansen & Yves Peirsman (eds.), *Recent Advances in Cognitive Sociolinguistics*, 111–137. Berlin/New York: Mouton de Gruyter.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast.In Joan Bybee & Paul Hopper (eds.) *Frequency and the emergence of linguistic structure*, 137-157. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Popper, Karl R. 1968 [1934]. *The logic of scientific discovery*. New York: Harper Torchbooks.

Pulman, S.G. 1993. *Word meaning and belief*. London/Canberra: Croom Helm.

Pijnenburg, Willy J.J. 1997. Vroegmiddelnederlands Woordenboek 1200–1300 (Dictionary of Early Middle Dutch: 1200-1300). In Karina H. van Dalen-Oskam et al. (Eds.), *Dictionaries of Medieval Germanic Languages: A Survey of Current Lexicographical Projects*, 3–10. Turnhout: Brepols.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org (last access 19.08.2011)

Rastier, François. 1999. Cognitive semantics and diachronic semantics: the values and evolution of classes. In Andreas Blank & Peter Koch (eds.), Historical semantics and cognition, 109–144. Berlin/New York: Mouton de Gruyter.

Robinson, Justina A. 2010. *Awesome* insights into semantic variation. In Dirk Geeraerts, Gitte Kristiansen & Yves Peirsman (eds.), *Recent*

Advances in Cognitive Sociolinguistics, 85–109. Berlin/New York: Mouton de Gruyter.

Rosch, Eleanor. 1975. Cognitive representation of semantic categories. *Journal of Experimental Psychology* 104 (3). 192–233.

Rosch, Eleanor. 1978. Principles of categorization. In Eleanor Rosch and Barbara.B. Lloyd (eds.), *Cognition and categorization*, 27–48. Hillsdale, NJ: Lawrence Erlbaum.

Rosch, Eleanor & Carolyn B. Mervis. 1975. Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology 7*, 573-605.

Rosch, Eleanor, Wayne D. Gray, David Johnson & Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8: 382–439.

Schmid, Hans-Jörg. 2010. Does frequency in text instantiate entrenchment? In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, 101–133. Berlin/New York: De Gruyter Mouton.

Schulte im Walde, Sabine. 2000. Clustering verbs semantically according to their alternation behaviour. In Proceeding of the 18th conference on Computational Linguistics – Volume 2, 747–753. Saarbrücken, Germany: Association for Computational Linguistics.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1): 97–123.

Shibatani, Masayoshi & Prashant Pardeshi. 2002. The causative continuum. In Masayoshi Shibatani (ed.), *The grammar of causation and interpersonal manipulation*, 85–126. Amsterdam/Philadelphia: John Benjamins.

Soares da Silva, Augusto. 2007. Verbs of letting: Some cognitive and historical aspects. In Nicole Delbecque & Bert Cornillie (eds.), *On interpreting construction schemas. From action and motion to transitivity and causality*, 171–200. Berlin/New York: Mouton de Gruyter.

Speelman, Dirk & Dirk Geeraerts. 2009. Causes for causatives: the case of Dutch 'doen' and 'laten'. In Ted Sanders and Eve Sweetser (eds.), *Causal categories in discourse and cognition*, 173–204. Berlin/New York: Mouton de Gruyter.

Stefanowitsch, Anatol. 2001. *Constructing causation: A Construction Grammar approach to analytic causatives*. Houston, TX: Rice University dissertation.

Stefanowitsch, Anatol. 2010. Empirical Cognitive Semantics: Some thoughts. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, 355–380. Berlin/New York: De Gruyter Mouton.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Stefanowitsch, Anatol & Stefan Th. Gries. 2008. Channel and constructional meaning: A collostructional case study. In: Kristiansen, Gitte & René Dirven (eds.), *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*, 129 – 152. Berlin: Mouton de Gruyter.

Storms, Gert, Paul De Boeck & Wim Ruts. 2000. Prototype and exemplar based information in natural language categories. *Journal of Memory and Language* 42. 51–73.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9 (307). http://www.biomedcentral.com/1471-2105/9/307 (last access 20 July 2011).

Stukker, Ninke. 2005. *Causality marking across levels of language structure*. University of Utrecht dissertation.

Szmrecsanyi, Benedikt. 2010. The English genitive alternation in a cognitive sociolinguistics perspective. In Dirk Geeraerts, Gitte Kristiansen and Yves Peirsman (eds.), *Advances in Cognitive Sociolinguistics*, 141–166. Mouton de Gruyter, Berlin/New York.

Tagliamonte, Sali A. & R. Harald Baayen. Submitted. Models, forests and trees in York English: Was/were variation as a case study for statistical practice.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.

Talmy, Leonard. 2007. Foreword. In Monica Gonzalez-Marquez, Irene Mittleberg, Seanna Coulson & Michael J. Spivey (eds.), *Methods in Cognitive Linguistics*, i-xxi. Amsterdam/Philadelphia: John Benjamins.

Taylor, John. 1989. *Linguistic categorization: Prototypes in linguistic theory.* Oxford: Clarendon.

Tummers, Jose, Kris Heylen & Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1(2). 225–261.

Vanpaemel, Wolf & Gert Storms. 2008. In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review* 15 (4). 732–749.

Verhagen, Arie. 1994a. Taalverandering en cultuurverandering: *doen* en *laten* sinds de 18e eeuw. [Language change and cultural change: *doen* and *laten* since the 18th century]. *Forum der Letteren* 35 (4). 256-271.

Verhagen, Arie. 1994b. Doen of laten: woordbetekenis of (ook) structuur? [Doen or laten: word meaning or (also) structure?] *Forum der Letteren* 35 (4). 277–281.

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: CSLI Publications.

Verhagen, Arie. 2007. English constructions from a Dutch perspective. In Mike Hannay and Gerard Steen (eds.), *Structural-Functional Studies in English Grammar*, 257–274. Amsterdam: John Benjamins.

Verhagen, Arie & Suzanne Kemmer. 1997. Interaction and Causation: Causative Constructions in Modern Standard Dutch. *Journal of Pragmatics* 27. 61–82.

Wierzbicka, Anna. 1985. *Lexicography and conceptual analysis*. Ann Arbor: Karoma.

Wierzbicka, Anna. 1988. *The semantics of grammar*. Amsterdam/Philadelphia: John Benjamins.

Wulff, Stefanie, Anatol Stefanowitsch & Stefan Th. Gries. 2007. Brutal Brits and persuasive Americans: Variety-specific meaning construction in the *into*-causative. In Günter Radden, Klaus-Michael Köpcke, Thomas Berg & Peter Siemund (eds.), *Aspects of meaning construction*, 265–281. Amsterdam/Philadelphia: John Benjamins.

Zeschel, Arne. 2010. Exemplars and analogy: Semantic extension in constructional networks. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, 201–219. Berlin/New York: De Gruyter Mouton.